Détection de sources quasi-ponctuelles dans des champs de données massifs

Céline Meillier

UNIVERSITÉ GRENOBLE ALPES

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE ALPES

Spécialité : **Signal, Image, Parole, Télécoms (SIPT)**

Arrêté ministériel : 7 août 2006

Présentée par

Céline MEILLIER

Thèse dirigée par **Olivier MICHEL**
et coencadrée par **Florent CHATELAIN**

préparée au sein **Gipsa-lab**
et de **l'école doctorale d'électronique, électrotechnique, automatique et traitement du signal (EEATS)**

Détection de sources quasi-ponctuelles dans des champs de données massifs

Thèse soutenue publiquement le **15 octobre 2015**,
devant le jury composé de :

Xavier DESCOMBES
Directeur de recherche INRIA, Président
David MARY
Professeur Université de Nice Sophia Antipolis, Rapporteur
Thomas RODET
Professeur ENS Cachan, Rapporteur
Roland BACON
Astronome, Directeur de recherche CRAL, Examinateur
Hervé CARFANTAN
Maître de conférence Université Paul Sabatier, Examinateur
Olivier MICHEL
Professeur Grenoble INP, Directeur de thèse
Florent CHATELAIN
Maître de conférence Grenoble INP, Encadrant de thèse
Hacheme AYASSO
Maître de conférence Université Joseph Fourier, Invité

Remerciements

Écrire les remerciements pour la version finale du manuscrit n'est définitivement pas la tâche la plus aisée de ces trois années de thèse. Écrire ces remerciements signifie la fin. La fin de trois ans de recherche sur un sujet passionnant qu'est la détection de galaxies, mais surtout la fin de trois années de vie au Gipsa durant lesquelles j'ai eu la chance de faire des rencontres exceptionnelles tant sur le plan scientifique que sur le plan humain. Alors avant de commencer à écrire des remerciements individuels (j'espère n'oublier personne!!) je souhaite surtout que l'achèvement de cette thèse ne soit pas la fin de toutes ces amitiés scientifiques et humaines que j'ai pu nouer durant trois ans.

Je vais commencer par remercier ceux sans qui cette thèse n'aurait pas pu avoir lieu : Florent et Olivier. Tout d'abord merci de m'avoir donné la chance d'expérimenter, en stage puis en thèse, le traitement du signal dans un domaine qui me passionnait depuis toujours, l'astronomie. Vous êtes tous les deux des encadrants hors pair et j'ai appris énormément de choses à vos côtés. Merci pour tout, pour toutes les discussions scientifiques (ou pas), pour avoir été disponibles et d'un soutien sans faille, même à deux semaines de la soutenance lorsqu'il a fallu refaire des calculs de dernière minute. Merci également à Hacheme pour tes conseils et le partage de tes connaissances en astronomie.

Cette thèse a été l'occasion de rencontrer les chercheurs, doctorants et postdoctorants du CRAL. Merci Roland pour la collaboration sur les données MUSE, pour avoir passé du temps à tester et t'approprier l'algorithme, j'espère que tu auras l'occasion de trouver des dizaines, voire des centaines de nouvelles galaxies. Merci à Laure d'avoir passer du temps à transformer mon code expérimental en un code livrable pour l'intégrer dans la suite logicielle associée aux données MUSE. Merci également à Johan, aux doctorants du CRAL que j'ai eu l'occasion de côtoyer, à Carole, Jean-Baptiste, Floriane que j'ai rencontrés au CRAL grâce à notre intérêt commun pour le traitement du signal appliqué à l'astronomie et pour les données MUSE.

Je tiens à remercier les membres du jury qui ont accepté d'évaluer mon travail de thèse. Un grand merci pour votre lecture attentive, vos analyses et retours constructifs.

Il est maintenant temps de remercier tous ceux qui ont contribué à faire de cette thèse une aventure humaine fantastique. Difficile de vous ranger dans des catégories pour faire des remerciements collectifs. Commençons d'abord par tous les co-bureaux : Cindy, Arnaud, Raluca, Pascal, Fahkri, Rafael, Fardin, merci pour les fous rires, le tableau des craquages, les Eusipcofolies, les lancers de schokobons à travers le bureau. Les partenaires de coinche le midi : Aude, Robin, Cindy, Quentin, Arnaud, Pascal, Manu, Romain, Edouard, Florian, Tim (celui avec des lunettes et deux yeux) et Tim (celui avec deux oreilles qu'on a obligé à jouer une ou deux fois), Raphael, Benoit (dit le chinois, qui n'est pas chinois), Maël, Taia, Lucas, Alexis, Rémy et bien d'autres encore... Les contributeurs au mur des cartes postales qui sont déjà cités dans les partenaires de coinches. Et voici en vrac tous ceux que je souhaite remercier (et re-remercier) en particulier :

— Cindy pour avoir été bien plus qu'une collègue de bureau. En trois ans nous sommes devenues amies, partenaires de coinche avec des jeux scandaleux, collaboratrices sans faille au sein du laboratoire, du gipsa-doc et de l'école doctorale et sources d'auto-motivation pour aller au sport le mardi soir et compenser les kinders buenos qu'on avalait avant d'y aller.
— Pascal, toujours prêt à partager une bouteille de Martini, toujours là quand on a besoin de toi, toujours prêt à organiser des vacances au soleil (avec une plage ! sinon ce ne sont pas des vacances).
— Arnaud pour les concours de paragraphes quand la motivation n'était pas à son maximum lors de la rédaction des articles et pour nous avoir déniché les meilleurs articles du web.
— Manu pour tes manues qui valent de l'or ("mais pourquoi tu me remercies ? tu ne me merci pas again").
— Raphael pour ton énergie débordante, ton soutien informatique qui défie n'importe quel service info, tes journées/soirées/weekends à optimiser mon code pour atteindre le facteur 2 (on a même dépassé le facteur 20).
— A&R, nos papi et mamie de compétition, toujours prêts à organiser une rando ou un après-midi jeux.
— Tim (avec des oreilles) pour tes questions sans queue ni tête et les sorties glaces, surtout en ces derniers mois de thèse (d'ailleurs est-ce que tu crois que le jour où tu écriras à ton tour les remerciements de ta thèse fera partie du top 3 des "meilleurs derniers jours de ..." et est-ce que ce top 3 fera partie du top 5 des meilleurs top 3 ?)
— Lucia, sans qui on serait tous atteints de phobie administrative, et pour ta bonne humeur, pour les chocolats qui apparaissent dans la cafétéria.

Merci également à tous les membres du DIS, en particulier à tous les membres de l'équipe CICS avec qui j'ai pu découvrir différents sujets de recherche y compris des tenseurs ou des télécommunications. Merci aussi aux membres des différentes équipes Gipsa-doc pour avoir animé la vie doctorante au labo avec autant d'entrain.

Pour terminer ces remerciements, un immense merci à ma famille, à mes parents surtout pour avoir toujours été là et m'avoir soutenue tout au long de cette thèse. Une pensée pour mes trois grand-parents, en particulier à Bernard et Liliane Meillier, qui j'en suis sure, auraient été fiers du chemin que j'ai parcouru.

Et enfin merci à toi, Quentin, pour être là pour moi au quotidien depuis de nombreuses années. Qui aurait cru qu'on se suivrait aussi longtemps, on en a fait du chemin depuis le lycée Victor Hugo !

Table des matières

Table des figures

Liste des tableaux

Liste des acronymes

BH Benjamini-Hochberg
DRS Data Reduction Software
EQM Erreur Quadratique Moyenne
FDP False Discovery Proportion
FDR False Discovery Rate
FSF Field Spread Function
FWER Family Wise Error Rate
FWHM Full Width at Half Maximum
HDFS Hubble Deep Field South
HST Hubble Space Telescope
IFU Integral Field Unit
i.i.d. (variables) indépendantes et identiquement distribuées
LSF Line Spread Function
MAP Maximum *a posteriori*
MCMC Markov Chain Monte Carlo
MUSE Multi Unit Spectroscopic Explorer
MV Maximum de vraisemblance
PSF Point Spread Function
RJMCMC Reversible Jump Markov Chain Monte Carlo
RSB Rapport Signal à Bruit
VLT Very Large Telescope

Introduction

Ce manuscrit présente les travaux réalisés dans le cadre de ma thèse intitulée *Détection de sources quasi-ponctuelles dans des champs de données massifs* menée au Gipsa-Lab. Bien que ces travaux puissent avoir un champ d'applications très vaste (la détection de sources est un problème classique du traitement du signal et de l'image), ils s'inscrivent dans un contexte applicatif bien particulier : la détection de galaxies lointaines dans les cubes de données hyperspectraux produite par le Multi Unit Spectroscopic Explorer (MUSE). Le chapitre 1 introduit en détail la problématique de la détection de galaxies dans le cadre du projet MUSE, nous proposons dans cette introduction de présenter brièvement les notions qui seront développées dans ce manuscrit.

Sources quasi-ponctuelles et champs de données massifs

L'observation d'un phénomène physique à travers un instrument optique donne lieu à un étalement de la réponse de la source observée dû à la fonction d'étalement du point (PSF pour *point spread function*). Lorsqu'une source est parfaitement ponctuelle ou de dimension inférieure à la limite de résolution, l'observation que nous ferons de la source est en réalité la PSF de l'instrument utilisé. Nous parlerons de sources quasi-ponctuelles si la dimension de la source est légèrement supérieure à la limite de résolution de l'instrument. Dans l'application astrophysique proposée dans ce manuscrit, les galaxies les plus lointaines observées avec l'instrument MUSE apparaissent comme un point dans l'espace et leur lumière est due à l'émission de photons à une longueur d'onde particulière. Avec la résolution de l'instrument MUSE, ces galaxies sont considérées comme quasi-ponctuelles.

Un champ de données désigne un ensemble cohérent de données, cela peut être une série temporelle, une image, un cube de données, etc. Le terme de données massives est souvent subjectif, nous l'utilisons dans ce manuscrit pour décrire des données dont la dimension est bien supérieure à la taille des signaux à détecter. Dans le cas du projet MUSE, les cubes hyperspectraux sont qualifiés de massifs comparés aux données hyperspectrales classiquement rencontrées dans la littérature. La dimension spatiale, 300×300 pixels, et le nombre de bandes spectrales, 3600 bandes, sont conséquents, les données hyperspectrales usuelles sont composées de quelques centaines de bandes. La dimension des données est à mettre en regard avec la taille des sources quasi-ponctuelles que nous cherchons à détecter.

Processus ponctuels marqués

Un processus ponctuel marqué est un processus aléatoire dont les réalisations sont des configurations d'objets. Nous modélisons la distribution spatiale des sources par une configuration de points distribués de façon aléatoire dans l'espace défini par les deux dimensions spatiales de l'image hyperspectrale. Des caractéristiques qui décrivent la forme des objets, leur intensité, leurs interactions avec les autres objets sont ajoutées à chaque point pour le transformer en objet. Le principe est de représenter un ensemble d'objets (qui peuvent être complexes de par leur forme, leurs caractéristiques spectrales, d'intensité, etc) par un processus objets simple ; le but n'est pas de reproduire parfaitement la complexité des données dans les marques des objets, mais plutôt

d'avoir une représentation de faible dimension (le nombre de marques caractérisant les objets est réduit) proche du phénomène physique. Dans le cas qui nous intéresse, au lieu de représenter les galaxies par une collection de pixels dont il faudrait décrire ensuite la distribution spatiale et spectrale, une galaxie sera modélisée par un point, son centre, et un profil d'intensité spatial (par exemple un profil gaussien en deux dimensions) associé à un support elliptique. Le spectre de la galaxie sera considéré comme une marque que nous estimerons *a posteriori*.

Estimation bayésienne

L'inférence bayésienne, qui repose sur le théorème de Bayes, conduit à définir un modèle paramétrique des observations disponibles et à exploiter les informations apportées par ces observations pour estimer les paramètres. Les observations sont vues comme la réalisation d'une variable aléatoire dont la loi est la fonction de vraisemblance. Les paramètres qui interviennent dans cette fonction sont aussi considérés comme des variables aléatoires distribuées selon une loi *a priori* qui quantifie l'incertitude que nous pouvons avoir sur les valeurs des paramètres. Cela permet d'obtenir une estimation robuste des paramètres du modèle directement à partir des données. Dans le cadre de la détection de galaxies dans les données MUSE, nous souhaitons mettre en place une méthode d'estimation du nombre, de la position, de la forme, du spectre, etc, des galaxies, robuste et qui s'appuie principalement sur l'information portée par les données.

Contrôle des erreurs

Dans tout processus de détection, dès lors que les observations sont bruitées, des erreurs vont se glisser parmi les détections. Etre en mesure de contrôler les erreurs permet d'interpréter la confiance que nous avons dans la détection.

Cahier de charges

La thèse comporte un aspect méthodologique important qui sera développé tout au long de ce manuscrit et un fort aspect algorithmique avec le développement d'un code de détection de galaxies dans les données MUSE qui a été transféré au consortium MUSE pour intégration dans la suite logicielle associée aux données. La conception logicielle sera très peu abordée dans ce manuscrit, cependant tous les résultats présentés ont été obtenus grâce au code développé durant la thèse.

Les objectifs méthodologiques sont :

- la prise en compte des très grandes dynamiques entre les galaxies pour détecter les galaxies les plus faibles sans être aveuglé par les galaxies les plus brillantes,
- la prise en compte des caractéristiques de l'instrument pour améliorer la localisation des objets et l'estimation de leurs caractéristiques spectrales,
- la mise en place d'*a priori* sur la configuration d'objets afin d'éviter les surdétections tout en permettant la détection de galaxies proches spatialement mais avec des spectres différents,
- l'optimisation de l'échantillonneur pour s'assurer de la convergence vers une solution quasi-optimale en un nombre fini d'itérations,
- le contrôle de la détection et l'extraction d'indices de confiance pour fournir aux astrophysiciens des outils supplémentaires pour mieux interpréter les résultats de la détection.

Dans ces travaux, nous proposons une méthode de détection en trois dimensions modélisant de manière parcimonieuse la configuration de galaxies tout en garantissant un certain contrôle des erreurs de détection. L'approche par processus ponctuel

marqué plongé dans un cadre entièrement Bayésien permet de proposer un algorithme robuste ne nécessitant aucun *a priori* sur les spectres des galaxies à rechercher.

Organisation du document

Le manuscrit est composé de quatre chapitres destinés à être lus de façon linéaire. Il faut noter cependant que les méthodes et algorithmes présentés dans le chapitre 3 peuvent être utilisés séparément dans différents contextes de détection. Un certain nombre d'annexes sont ajoutées en fin de manuscrit afin de détailler certaines notions fondamentales utilisées au cours du manuscrit (processus ponctuels marqués, tests multiples), ou encore pour introduire certains modèles mathématiques (représentation d'un objet elliptique pour représenter les galaxies, matrice de filtrage).

Nous introduisons en quelques pages les notations utilisées pour décrire les données hyperspectrales tout au long de ce manuscrit. En effet, les données hyperspectrales sont composées de différentes informations sous différents formats : le cube de données dans sa globalité, des images, des spectres, des pixels. Chacun de ces formats possède sa typographie, reportée dans la partie intitulée *Notations*.

Le chapitre 1, intitulé *Détection de galaxies lointaines dans les images hyperspectrales MUSE*, nous permet d'introduire le contexte du projet MUSE : de la mise en oeuvre de l'instrument à la problématique des galaxies lointaines. Différentes méthodes de la littérature pour la détection de sources en imagerie hyperspectrale et en astronomie sont détaillées dans ce chapitre. Nous présenterons également le jeu de données synthétiques sur lequel sont basées les analyses de performance de la méthode de détection proposée dans le chapitre 2 et les prétraitements décrits dans le chapitre 3 ainsi que le jeu de données réelles sur lequel nous appliquerons la méthode proposée. Nous introduirons enfin dans ce chapitre le cahier des charges établi au début de la thèse pour l'élaboration de la méthode de détection.

Le chapitre 2, intitulé *La méthode de détection*, traite de la modélisation des données et de la configuration de galaxies que nous souhaitons détecter. Le modèle est formulé dans un cadre entièrement bayésien qui permet une estimation robuste à partir des données des paramètres du modèle. La méthode de détection basée sur un algorithme itératif de type échantillonnage par méthode de Monte Carlo par chaîne de Markov à sauts réversibles est décrite, ses performances, ses avantages et ses limitations sont également étudiés dans ce chapitre.

Le chapitre 3, intitulé *Prétraitements des données*, peut se lire indépendamment du reste du manuscrit. Nous introduisons la nécessité de prétraiter les données, d'une part pour guider la méthode de détection dans sa recherche d'objets dans les zones les plus probables, d'autre part, pour fournir un moyen de contrôler les erreurs de détection. Pour répondre à ce besoin, nous introduirons différentes méthodes de la littérature basées sur la formulation de tests multiples et le contrôle des erreurs pour un ensemble de tests. L'étude de ces méthodes nous a permis de formuler deux types de tests, et donc deux types de contrôle des erreurs différents, que nous appliquerons sur les données MUSE.

Le chapitre 4, intitulé *Application aux données réelles*, présente une analyse complète, du prétraitement des données à l'analyse du catalogue d'objets détectés, d'un jeu de données réelles. Ces données sont particulièrement bien connues des astrophysiciens puisqu'elles ont déjà été observées à l'aide d'autres instruments. Travailler avec ces données fournit l'avantage de pouvoir comparer les résultats de notre méthode de détection sur les données MUSE avec un catalogue de sources identifiées sur les observations précédemment réalisées. Nous attacherons une attention particulière à l'analyse des objets détectés, aux objets manqués et aux fausses détections. Nous

analyserons les points positifs et les limitations de l'algorithme au vu des résultats obtenus sur le jeu de données réelles.

Publications de l'auteur

Article de journal

- Meillier et al. [2015a] : C. Meillier, F. Chatelain, O. Michel, and H. Ayasso. Nonparametric bayesian extraction of object configurations in massive data. *Transactions on Signal Processing, IEEE* , 63(8) :1911-1924, 2015.

Cet article de journal paru dans *Transactions on Signal Processing, IEEE* détaille le principe de la méthode de détection de sources quasi-ponctuelles en insistant notamment sur la description par processus ponctuel marqué, présenté en détail dans l'annexe B, et sur le cadre bayésien dans lequel se déroule l'estimation des paramètres, présenté dans le chapitre 2.

- Meillier et al. [2016] : C. Meillier, F. Chatelain, O. Michel, R. Bacon, L. Piqueras, R. Bacher and H. Ayasso. SELFI : an object based, Bayesian method for faint emission line source detection in MUSE deep field datacubes. Pour publication dans *Astronomy and Astrophysics*.

Cet article de journal soumis à *Astronomy and Astrophysics* résume le principe de la méthode de détection de sources quasi-ponctuelles appliquée au cas de la détection des galaxies lointaines de faible intensité observées par l'instrument MUSE. L'analyse des résultats de la détection sur le champ profond HDFS acquis par MUSE en 2014 est présentée dans ce papier.

Conférences internationales

- Meillier et al. [2014] : C. Meillier, F. Chatelain, O. Michel, and H. Ayasso. Non-parametric bayesian framework for detection of object configurations with large intensity dynamics in highly noisy hyperspectral data. *In Proc. IEEE International Conference Acoustics, Speech and Signal Processing (ICASSP)*, pages 1886-1890, 2014.

Cet article de conférence présenté à ICASSP résume la méthode de détection présentée au chapitre 2 ainsi que le prétraitement basé sur le max-test présenté dans le chapitre 3 de ce manuscrit. Un exemple de détection est donné sur le cube de données synthétiques que nous allons également exploiter dans ce manuscrit.

- Meillier et al. [2015b] : C. Meillier, F. Chatelain, O. Michel, and H. Ayasso. Error control for the detection of rare and weak signatures in massive data. *In Proc. European Signal Processing Conference (EUSIPCO)*, 2015.

Cet article de conférence présenté à EUSIPCO introduit un concept statistique, le contrôle des fausses découvertes, qu'il est possible d'appliquer à un outil emblématique du traitement du signal, le filtrage adapté, à condition que les réponses du filtre et des sources soient positives. Cela correspond aux résultats présentés dans le chapitre 3 de ce manuscrit.

Conférences nationales

- Meillier et al. [2015c] : C. Meillier, F. Chatelain, O. Michel, and H. Ayasso. Contrôle des erreurs pour la détection d'événements rares et faibles dans des champs de données massifs. *GRETSI*, 2015.

Cet article de conférence présenté au GRETSI propose à la communauté française du traitement du signal des résultats similaires à ceux présentés dans Meillier et al. [2015b].

Notations

Dans ce manuscrit, nous allons manipuler des données sous forme de cubes hyperspectraux. Selon le contexte, nous aurons besoin d'exprimer certaines quantités dans un espace en trois dimensions, sous forme vectorisée dans un espace à une dimension ou encore à l'aide d'images extraites du cube de données. Le lecteur pourra se référer à cette page d'introduction aux notations tout au long de sa lecture du manuscrit.

Remarques générales

Nous ferons référence tout au long du manuscrit aux différents élements du cube de données MUSE avec les notations suivantes :

- $\mathbf{Y}$ et toutes ses variantes ($\boldsymbol{y}$, $\mathbf{Y}^{(v)}$, etc) désigneront le cube de données MUSE sous différentes formes,
- (p, q) est utilisé à la place des traditionnelles coordonnées cartésiennes (x, y) afin d'éviter les confusions avec d'autres notations utilisées dans le manuscrit. Dans un tableau numérique, p indique l'indice de la ligne considérée et q l'indice de la colonne.
- λ désigne aussi bien la longueur d'onde exprimée en nanomètre (nm) ou en Angström (Ang) [1] que l'indice du tableau dans lequel sont stockées les données MUSE correspondant à la longueur d'onde exprimée en nanomètre ou en Angtröm.

Nous utiliserons de manière implicite les coordonnées (p, q, λ) pour désigner aussi bien une position continue dans la scène observée qu'une position numérique dans le cube (p, q et λ sont alors des entiers). Nous considérons alors que la scène observée appartenant au monde réel, et donc continue, est projetée sur une grille discrète de pixels.

Notations dans un espace à trois dimensions

Les données hyperspectrales MUSE sont stockées numériquement dans un tableau de dimension trois représentant les deux dimensions spatiales et la dimension spectrale. La figure 1 introduit les notations utilisées lorsque nous travaillons avec le cube de dimension trois. En dimension trois, nous gardons la notation informatique pour désigner les différents objets que nous allons manipuler, ainsi :

- le cube est noté $\mathbf{Y}$,
- un pixel du cube à la position (p, q, λ) est noté $\mathbf{Y}(p, q, \lambda)$,
- l'image correspondant à la lumière enregistrée à la longueur d'onde λ est notée $\mathbf{Y}(\cdot, \cdot, \lambda)$,
- le spectre correspondant à l'intensité sur toutes les longueurs d'onde de l'instrument pour la position spatiale (p, q) est noté $\mathbf{Y}(p, q, \cdot)$.

1. 1nm = 10 Ang.

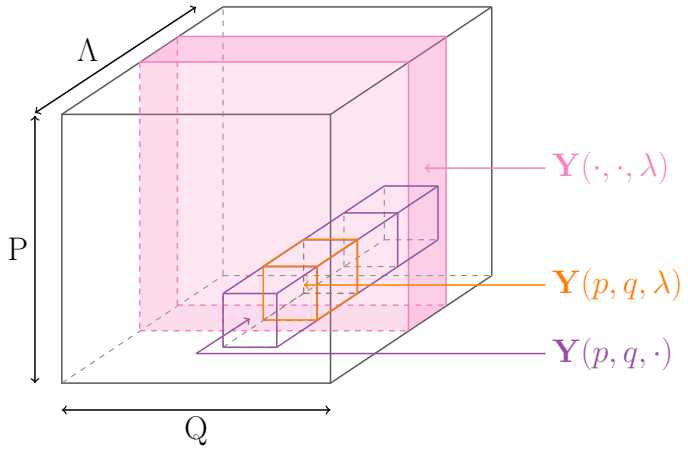


Figure 1 – Notations exprimées dans un espace à trois dimensions pour le cube $\mathbf{Y}$, une image $\mathbf{Y}(\cdot,\cdot,\lambda)$ correspondant aux observations à la longueur d'onde λ fixée, un spectre $\mathbf{Y}(p,q,\cdot)$ et enfin un pixel $\mathbf{Y}(p,q,\lambda)$.

Notations dans un espace vectorisé (une dimension)

Dans les chapitres 2, 3 et 4, nous introduirons des méthodes de traitement sur les données qui nécessitent parfois de les mettre sous forme de vecteurs.

Image

La vectorisation d'une image de taille $P \times Q$ est présentée sur la figure 2. Afin de vectoriser une image, les colonnes de l'image sont alignées les unes en dessous des autres pour former un vecteur colonne de taille $(P \times Q) \times 1$. Cette notation sera notamment utilisée dans le chapitre 2 pour définir le modèle d'observation des données.

Spectre

Lorsqu'un spectre est extrait du cube de données pour être traité séparément, la notation $\mathbf{Y}(p,q,\cdot)$ est allégée en $\mathbf{y}_r(\cdot)$ où r est l'équivalent dans l'espace vectorisé de la position spatiale (p,q). Les modifications de notations sont représentées sur la figure 3. La notation allégée sera utilisée dans le chapitre 3 dans lequels sont présentés des tests appliqués aux spectres du cube.

Cube

La vectorisation du cube ne sera pas représentée graphiquement pour des raisons d'économie de place dans le manuscrit. Afin de former le vecteur $\mathbf{Y}^{(v)}$ de taille $(P \times Q \times \Lambda) \times 1$ à partir du cube $\mathbf{Y}$, les images sont tout d'abord vectorisées selon la méthode représentée sur la figure 2, puis elles sont alignées les unes en dessous des autres.

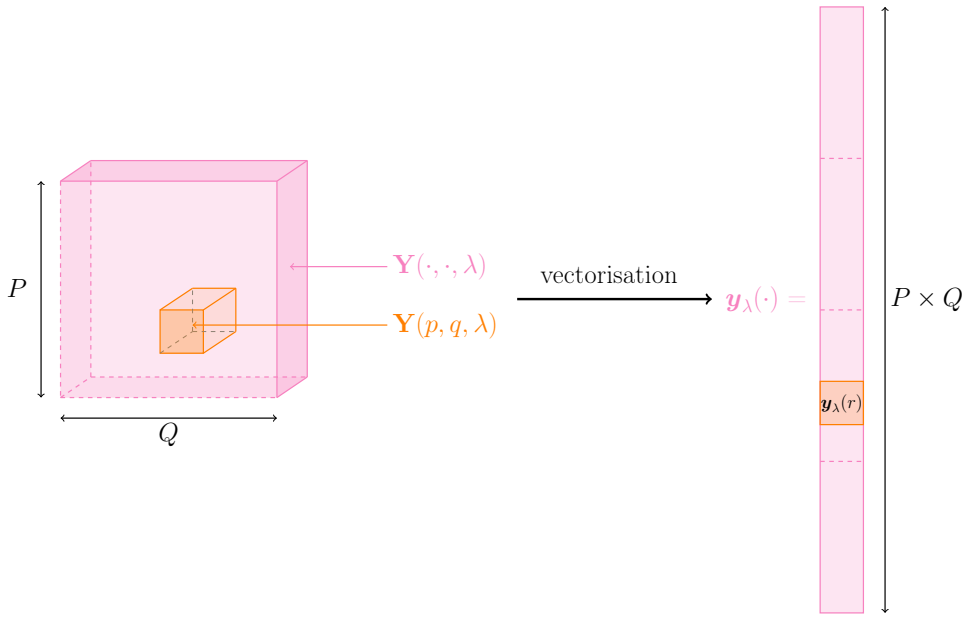


Figure 2 – La vectorisation d'une image entraîne une modification des notations : l'image $\mathbf{Y}(\cdot,\cdot,\lambda) \in \mathbb{R}^{P\times Q}$ devient $\boldsymbol{y}_\lambda(\cdot) \in \mathbb{R}^{M\times 1}$ où $M = P \times Q$, et le pixel $\mathbf{Y}(p,q,\lambda)$ devient alors $\boldsymbol{y}_\lambda(r)$ avec $r \equiv (p,q)$.

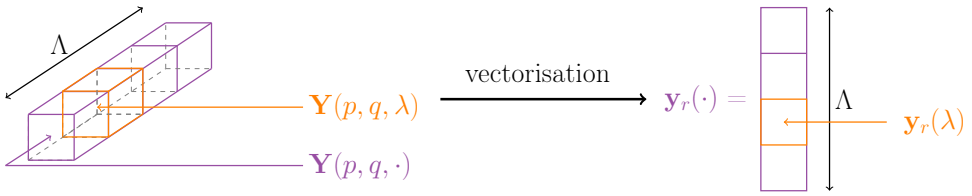


Figure 3 – Modifications des notations pour le traitement des spectres.

Récapitulatif

Le tableau 1 fait le récapitulatif des différentes notations introduites précédemment pour représenter les données.

Objets	3D	1D	Remarques
Cube	$\mathbf{Y}$	$\mathbf{Y}^{(v)}$	L'indice (v) signifie *vectorisé*
Image	$\mathbf{Y}(\cdot,\cdot,\lambda)$	$\boldsymbol{y}_\lambda$	L'indiçage par lettre grecque désigne une longueur d'onde.
Spectre	$\mathbf{Y}(p,q,\cdot)$	$\mathbf{y}_r(\cdot)$	$r \equiv (p,q)$ avec $r = (q-1) \times P + p$
Pixel	$\mathbf{Y}(p,q,\lambda)$	$\boldsymbol{y}_\lambda(r)$	$r \equiv (p,q)$ pixel dans une image vectorisée
		$\mathbf{y}_r(\lambda)$	$r \equiv (p,q)$ pixel dans un spectre
		$\mathbf{Y}^{(v)}(r)$	$r \equiv (p,q,\lambda)$ avec $r = (\lambda-1) \times P \times Q + (q-1) \times P + p$

Tableau 1 – Récapitulatif des notations dans l'espace de dimension trois (3D) et l'espace vectorisé (1D).

Chapitre 1

Détection de galaxies lointaines dans les images hyperspectrales MUSE

Sommaire

Ce premier chapitre a pour but d'introduire le contexte général de détection de galaxies lointaines dans les données MUSE. Pour ce faire, nous présenterons brièvement l'instrument MUSE, ses spécificités techniques et le grand challenge scientifique pour lequel il a été construit : la détection de galaxies lointaines à l'aide de l'une de leurs caractéristiques spectrales, la raie d'émission Lyman α. Nous verrons que la communauté astrophysique a déjà proposé des méthodes de détection de sources dans des images en deux dimensions, et très récemment dans des cubes de données similaires à ceux produits par MUSE. Nous nous pencherons également vers les stratégies de détection développées dans la littérature hyperspectrale. Nous verrons cependant qu'aucune de ces méthodes ne remplit le cahier des charges que nous nous sommes fixé pour traiter les

données MUSE. Dans la partie 1.4, les deux grands jeux de données utilisés durant cette thèse seront présentés. Enfin dans la dernière partie de ce chapitre, nous présenterons les hypothèses adoptées pour la modélisation numérique de l'influence de l'instrument sur les observations, la modélisation du bruit et des galaxies présentes dans les données MUSE.

1.1 Le projet MUSE

Bien qu'applicable à une variété de domaines différents, la méthode de détection de sources quasi-ponctuelles dans des champs de données massifs qui est développée dans ce manuscrit est tout d'abord une réponse apportée au problème de détection de galaxies lointaines dans les données hyperspectrales MUSE. Nous présenterons dans ce paragraphe le projet MUSE, les données produites par l'instrument et la problématique de détection.

1.1.1 Le consortium

L'instrument MUSE (Multi Unit Spectroscopic Explorer) est issu de la création du consortium du même nom, piloté par le Centre de Recherche Astrophysique de Lyon (CRAL), sous la direction de Roland Bacon. Ce consortium regroupe plusieurs centres de recherche européens : l'*European Southern Observatory* (ESO), l'Institut de Recherche et d'Astrophysique et de Planétologie (IRAP), l'Observatoire de Leiden, l'Institut d'Astrophysique de Göttingen, l'École polytechnique fédérale de Zurich (ETH) et l'Institut Leibniz d'Astrophysique de Potsdam. Le consortium regroupe un ensemble de spécialités nécessaires à l'élaboration de l'instrument, l'exploitation et l'analyse des données produites : optique, mécanique, électronique, cryogénie, traitement du signal, management, astrophysique instrumentale et théorique.

1.1.2 L'instrument

MUSE est l'un des quatre instruments de seconde génération validés par l'ESO afin de renouveler l'instrumentation du *Very Large Telescope* (VLT) qui est représenté sur la figure 1.1. Le VLT est constitué de quatre télescopes primaires (appelés UT pour *Unit Telescope* en anglais) et quatre télescopes secondaires. Les télescopes primaires peuvent fonctionner simultanément en mode interférométrie pour former le Very Large Telescope Interferometer (VLTI) ou séparément en s'associant à un instrument de type imageur CCD grand champ, caméra ou spectrographe. Le VLT constitue l'une des plus grandes installations permettant l'observation terrestre dans le domaine du visible et de l'infrarouge. La figure 1.2 montre l'instrument MUSE installé au foyer

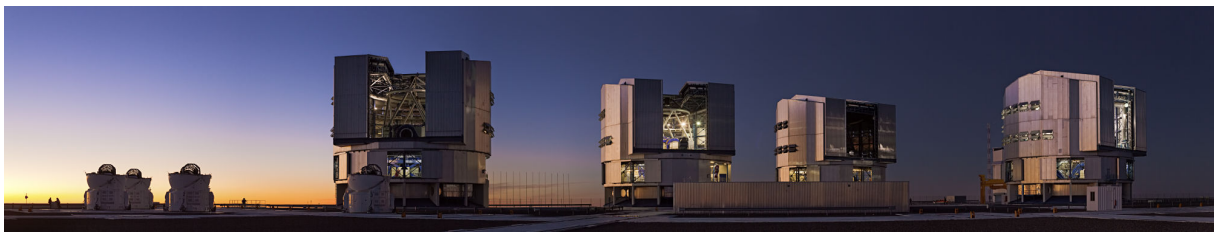


Figure 1.1 – Photographie du VLT à Paranal (Chili) avec ses quatre télescopes primaires et ses quatre télescopes secondaires. Photo disponible sur le site de l'ESO, crédit : ESO/B. Tafreshi (twanight.org).

Nasmyth de l'UT4, le quatrième télescope primaire du VLT. Les 24 spectrographes constituant MUSE bénéficient ainsi de la puissance de l'UT4 et de son miroir primaire de 8,2 mètres de

diamètre. MUSE est en opération au VLT depuis le 31 janvier 2014, date de sa première lumière.

FIGURE 1.2 – Photographie de l'instrument MUSE (à gauche) installé au foyer de l'UT4 (au centre). Photo disponible sur le site de l'ESO, crédit : Eric Le Roux/University Claude Bernard Lyon 1/CNRS/ESO.

MUSE est un spectrographe intégral de champ (ou spectrographe 3D), c'est-à-dire qu'il fournit non seulement une image du champ observé, avec un échantillonnage spatiale de 0,2 arcseconde, mais aussi un spectre pour chacun des pixels de cette image avec un échantillonnage spectral de 0,125 nm. L'analyse du spectre des objets observés permet de les localiser en distance relative par rapport à la Terre grâce à l'effet Doppler-Fizeau. En effet les raies d'émission des éléments chimiques présents dans les galaxies subissent un décalage dans le rouge qui est proportionnel à leur distance. L'analyse de ce décalage (appelé aussi *redshift*) permet de déduire la distance de l'objet observé. Ainsi MUSE produit une carte du ciel en trois dimensions contrairement aux imageurs classiques qui travaillent soit à une longueur d'onde particulière soit en moyennant le signal observé sur une large plage de longueurs d'onde.

L'instrument MUSE est composé de 24 modules identiques qui contiennent chacun un spectrographe. Un module est appelé IFU pour *Integral Field Unit*, la conception d'un IFU est décrite dans l'article de Laurent et al. [2006]. Le trajet de la lumière entre son arrivée sur le miroir primaire de l'UT4 et la production du cube de données hyperspectrales est complexe. C'est aussi ce qui fait l'originalité de l'instrument MUSE. La lumière est réfléchie par le miroir primaire de l'UT4, elle est ensuite redirigée par un jeu de miroirs secondaires sur un dérotateur de champ qui compense l'effet de la rotation de la Terre sur les observations. La lumière est alors envoyée sur un découpeur de champ (*slicer* en anglais, voir Laurent et al. [2014]) qui sépare le champ observé en 24 sous-champs qui sont envoyés chacun sur un des 24 IFU de l'instrument MUSE. Dans chaque IFU, il y a un nouveau découpeur de champ qui découpe le sous-champ en 48 tranches qui sont ensuite envoyées dans le spectrographe. Ce spectrographe sépare la lumière en longueurs d'onde (bandes étroites de 0,125 nm de largeur) et le résultat est enregistré par un capteur CDD de 4000×4000 pixels. Les données ainsi collectées sont finalement réorganisées sous la forme d'un cube hyperspectral à l'aide du système de réduction de données (DRS pour *Data Reduction Software* en anglais) décrit dans les travaux de Weilbacher et al. [2012]. La figure 1.3 résume le trajet de la lumière.

La complexité des données MUSE est liée à son aspect hyperspectral massif (plus de 90000 spectres composés de plus de 3600 longueurs d'onde). Un cube de données couvre ainsi un grand domaine spectral, $\lambda \in [460\text{nm}, 930\text{nm}]$ échantillonné tous les 0,125nm, englobant le domaine visible et le proche infrarouge. La dimension des données MUSE est approximativement de 300×300 pixels pour les dimensions spatiales (soit un champ couvrant $1' \times 1'$ (minute d'arc) avec une résolution spatiale de 0.2 arcsec/pixel). Soit plus de 324 millions de pixels à analyser.

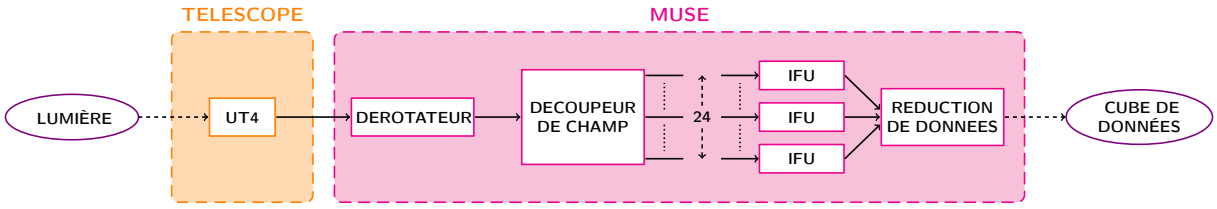


FIGURE 1.3 – Les différentes étapes du trajet de la lumière depuis son arrivée sur le miroir primaire de l'UT4 jusqu'à la construction du cube de données à l'aide du logiciel de réduction de données (DRS).

Une représentation symbolique d'un cube de données MUSE de taille standard est donnée par la figure 1.4.

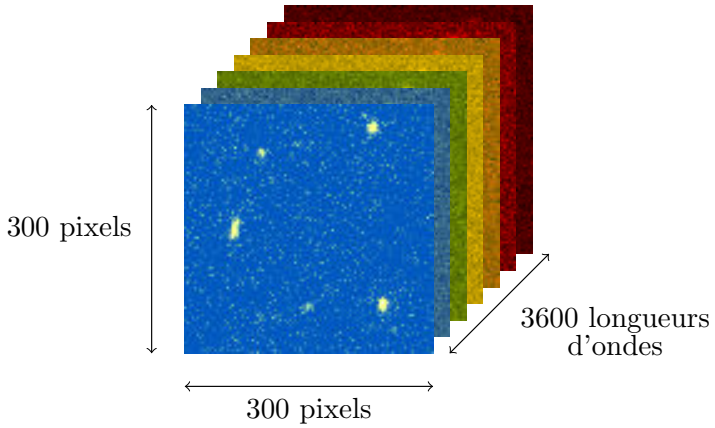


FIGURE 1.4 – Représentation d'un cube de données hyperspectral MUSE.

1.1.3 A la recherche des galaxies lointaines

L'un des axes scientifiques majeurs du projet MUSE est l'étude de la formation des galaxies. Bien que de nombreuses études soient consacrées à la modélisation numérique de la formation des galaxies, afin de valider un modèle, il faudrait pouvoir observer de très jeunes galaxies.

Pour répondre à ce problème il faut se référer au principe cosmologique, mis en évidence par Einstein en 1917 pour expliquer les équations décrivant l'Univers dans sa globalité à travers la théorie de la relativité générale. E. Milne l'énoncera sous le nom de principe cosmologique en 1930. Le principe cosmologique est l'hypothèse, généralement adoptée pour écrire un modèle cosmologique, que l'Univers est spatialement homogène et isotrope, *i.e.* l'observation que l'on fait de l'Univers (à grande échelle) ne dépend pas de la position de l'observateur. Ainsi pour observer la jeunesse de galaxies telles que la Voie Lactée ou d'autres galaxies proches de la nôtre, il suffit d'observer des galaxies lointaines, car regarder loin dans l'espace, c'est regarder loin dans le passé. La lumière que l'on perçoit aujourd'hui d'une galaxie située à 10 milliards d'années-lumière de la Terre a été émise il y a 10 milliards d'années. Et d'après le principe cosmologique, le passé de ces galaxies est similaire au passé de notre galaxie.

Observer des galaxies situées à plus de 10 milliards d'années-lumière n'est pas une tâche facile. En premier lieu, la lumière émise par ces galaxies lointaines est tellement faible, que même avec les grands télescopes terrestres (VLT) ou spatiaux (Hubble) il faut combiner plusieurs dizaines

d'heures de pose pour obtenir suffisamment de photons sur le détecteur et pouvoir ainsi les détecter. La taille perçue de ces galaxies lointaines sera d'autant réduite qu'elles seront éloignées de la Terre. Plus leur intensité sera faible et plus l'occupation spatiale sera petite, plus elles seront difficiles à détecter. A cela viennent s'ajouter les perturbations atmosphériques dans le cas des instruments terrestres comme l'ensemble VLT + MUSE. En effet, la lumière émise par une galaxie lointaine doit traverser l'atmosphère de la Terre qui est constituée de gaz et d'aérosols qui vont engendrer des effets d'absorption et de diffusion du signal lumineux.

A la différence de notre galaxie qui ne produit plus guère que trois étoiles de la taille de notre Soleil par an, les très jeunes galaxies sont un vivier d'étoiles en formation. En effet, les jeunes galaxies contiennent une très grande quantité d'hydrogène, élément primordial à la formation des étoiles. C'est cette caractéristique qui va permettre à l'instrument MUSE d'observer les galaxies lointaines malgré tous les obstacles situés sur le trajet de la lumière énoncés précédemment. Les jeunes galaxies n'émettent pas de manière uniforme sur toute la gamme de longueurs d'onde ; bien au contraire, étant composées essentiellement d'étoiles en formation, la quasi totalité de l'énergie lumineuse se trouve sur des raies d'émission de l'atome d'hydrogène. Les trois premières séries de transitions électroniques de l'atome d'hydrogène sont représentées sur la figure 1.5. La série de Lyman correspond aux transitions des états excités $n \geqslant 2$ vers son état fondamental $n = 1$, avec n le nombre quantique principal, la série de Balmer correspond aux transitions des états $n \geqslant 3$ vers l'état $n = 2$ et la série de Paschen correspond aux transitions des états $n \geqslant 4$ vers l'état $n = 3$. A chacune de ces transitions, on peut associer la longueur d'onde correspondant au niveau d'énergie du photon libéré lors de la transition. Ainsi pour une transition de l'état $n = 2$ vers l'état $n = 1$, on observe une raie d'émission appelée raie Lyman alpha (notée Lyα) à la longueur d'onde $\lambda_{Ly\alpha} = 121,567\text{nm}$). Cette raie correspond à la raie d'intensité la plus forte parmi les raies de la série de Lyman dans le spectre d'émission.

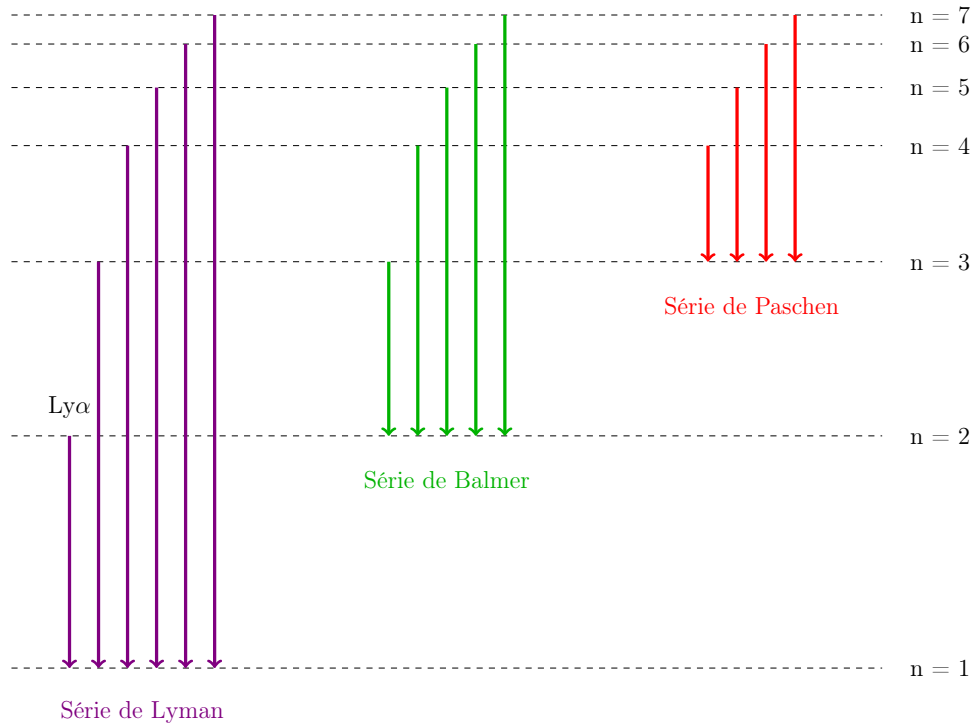


FIGURE 1.5 – Représentation des transitions électroniques de l'atome d'hydrogène pour les 3 premières séries, la série de Lyman (ultraviolet), la série de Balmer (visible) et la série de Paschen (infrarouge).

Un fort taux d'émission dans la raie Lyα est caractéristique des galaxies composées majoritairement d'étoiles jeunes ou en formation. La raie Lyα est donc un marqueur essentiel pour traquer les galaxies jeunes. Afin d'améliorer significativement le rapport signal à bruit des galaxies lointaines, il faut donc restreindre l'observation aux longueurs d'onde situées autour de la raie Lyα, une grande partie de la lumière émise sera concentrée autour de cette longueur d'onde. C'est ce que font les imageurs classiques qui, au lieu de moyenner la lumière sur la totalité du domaine spectral, utilisent un filtre sélectif en longueur d'onde. Cependant, dans le cas d'un champ d'observation profond comme ceux qui sont produits par MUSE, les galaxies lointaines ne sont pas toutes situées à la même distance. En vertu du principe de l'effet Doppler-Fizeau, les raies d'émissions de l'atome d'hydrogène verront leur longueur d'onde translatée dans le rouge proportionnellement à la vitesse radiale relative v_r de la galaxie le long de la ligne de visée selon la relation suivante :

$$z = \frac{v_r}{c}$$

où z est le décalage dans le rouge (appelé aussi *redshift* en anglais), et c est la célérité de la lumière dans le vide. À faible z, la loi de Hubble permet de relier la vitesse d'éloignement v_r à la distance d de la galaxie par l'équation :

$$v_r = H_0 d$$

avec H_0 la constante de Hubble. À grand z, la relation devient :

$$v_r = \frac{(z+1)^2 - 1}{(z+1)^2 + 1} c.$$

Le redshift peut également s'écrire en fonction de la longueur d'onde émise et la longueur d'onde observée :

$$1 + z = \frac{\lambda_{obs}}{\lambda_0}$$

où λ_{obs} est la longueur d'onde de la raie dans le spectre observé et λ_0 est la longueur d'onde correspondant à la transition électronique mesurée dans le référentiel de l'observateur. On peut ainsi mesurer la distance d'une galaxie observée grâce à la raie Lyα et son décalage dans le rouge observé dans le spectre. La puissance de l'instrument MUSE provient justement de sa capacité à observer un large champ sur une grande gamme de longueurs d'onde avec une résolution spectrale très fine. Ainsi pour une gamme de longueurs d'onde $\lambda \in [460\text{nm}, 930\text{nm}]$, les galaxies de type Lyα observées par MUSE peuvent présenter un décalage dans le rouge $z \in [2.78, 6.65]$, ce qui correspond à des galaxies dont l'âge varie de 11.35 milliards d'années à 13 milliards d'années.

Détecter des galaxies lointaines dans les données MUSE revient donc à chercher des sources spatialement peu étendues, avec un spectre constitué essentiellement d'une raie d'émission dans les quelques 324 millions de pixels qui constituent le cube de données hyperspectrales.

1.1.4 La détection de galaxies dans les données MUSE

Les données MUSE sont des observations en champ profond (quelques dizaines d'heures de pose). Elles contiennent de nombreuses galaxies lointaines que l'on souhaite détecter afin de permettre aux astrophysiciens de les étudier plus en détail pour comprendre la formation des galaxies dans l'histoire de l'Univers. Cependant la détection de galaxies lointaines fait face à différents obstacles que nous avons tenté de résoudre dans cette thèse :

— le nombre de galaxies observées est inconnu, tout comme leur position, leur forme ou leur spectre. Nous ne disposons d'**aucune connaissance *a priori* sur la configuration** pour réaliser la détection des galaxies observées dans le cube de données,

— il faut aussi tenir compte du fait que les galaxies lointaines ont une faible extension spatiale, du fait de leur distance, et que leur spectre est composé essentiellement d'une raie d'émission (la raie Lyα) qui s'étend sur quelques nanomètres, soit une signature présente sur quelques bandes spectrales. Le domaine d'existence de ces galaxies (quelques pixels adjacents dans les trois dimensions) est à comparer avec la dimension des données (plus de 324 millions de pixels). On peut parler ici de **détection de sources quasi-ponctuelles en trois dimensions dans des données massives**,
— parmi toutes les sources observées, se trouvent également quelques étoiles et des galaxies proches qui sont très brillantes, spatialement étendues et possèdent éventuellement une composante spectrale continue de forte amplitude. Il faudra donc **prendre en compte les grandes dynamiques entre les amplitudes des sources observées** pour réussir à détecter les plus petites et les moins brillantes d'entre elles,
— les données sont très bruitées, dans le cas de la détection des galaxies de très faible intensité, une bonne **modélisation du bruit** ainsi que certains **prétraitements des données** sont nécessaires afin de maximiser les chances de détection,
— **le profil spectral de la raie Lyα est très variable**, il peut comporter une raie simple ou une raie double, présenter un profil très asymétrique et la largeur de la raie n'est pas fixe (voir par exemple les observations réalisées par Tapken et al. [2007] et les travaux de thèse de Garel [2011]). Il sera donc impossible de travailler à partir d'un gabarit précis pour détecter les sources comportant une raie Lyα dans le cube.

1.2 La détection de sources en imagerie hyperspectrale

Puisque les données MUSE sont avant tout des images hyperspectrales, il est intéressant de se pencher sur les méthodes de détection de sources développées dans la littérature hyperspectrale. Historiquement les données hyperspectrales sont issus d'applications de télédétection. Le premier modèle de données utilisé, le modèle de mélange linéaire des signatures spectrales des matériaux observés sur une scène, a donné lieu à l'élaboration d'un grand nombre de méthodes visant à extraire des données ces signatures spectrales et leurs fractions dans chaque pixel de l'image. Le principe général de ces méthodes sera présenté dans le paragraphe 1.2.1. Nous nous intéresserons ensuite aux méthodes de détection de cibles, c'est-à-dire d'une signature spectrale particulière, dans le paragraphe 1.2.2. Nous verrons enfin dans le paragraphe 1.2.3 que parmi toutes ces approches pixelliques, il existe une méthode de détection d'objets.

1.2.1 Méthodes d'extraction des pôles de mélange et d'estimation des abondances

Dans la littérature hyperspectrale, la majorité des approches développées pour la détection de sources sont orientées vers l'estimation de la signature spectrale des composants de la scène observée et de la fraction de chacune de ces composantes à chaque position (x, y) spatiale du cube de données. Les signatures des composants purs sont appelées *pôles de mélange* ou *endmembers* en anglais, et la fraction de ces pôles de mélange dans l'image, l'*abondance*. Les problèmes de démélange dans le visible et le proche infrarouge reposent en général sur un modèle de mélange linéaire :

$$\mathbf{Y} = \mathbf{SA} + \boldsymbol{\epsilon} \tag{1.1}$$

où $\mathbf{Y} \in \mathbb{R}^{\Lambda \times \mathrm{M}}$ est l'image hyperspectrale vectorisée avec M le nombre de pixels par image et Λ le nombre de bandes spectrales. La matrice $\mathbf{S} \in \mathbb{R}^{\Lambda \times \mathrm{L}}$ est une matrice de L composantes spectrales pures. Le nombre L de composantes de $\mathbf{S}$ peut être égal au nombre de pôles de mélange de l'image s'il est connu ou estimé directement à partir des données, il peut être plus grand si $\mathbf{S}$ est un dictionnaire regroupant les signatures spectrales (connues, mesurées) d'un grand nombre de

matériaux différents. La matrice $\mathbf{A} \in \mathbb{R}^{\mathrm{L}\times\mathrm{M}}$ contient les abondances de chacune des composantes de la matrice $\mathbf{S}$ en chacun des pixels de l'image. La matrice $\boldsymbol{\epsilon} \in \mathbb{R}^{\Lambda\times M}$ représente les erreurs de reconstruction et le bruit de mesure. Les méthodes de démélange faisant appel à des dictionnaires de signatures connues ne sont pas envisageables pour traiter les données MUSE pour plusieurs raisons :

- les spectres des galaxies sont *a priori* inconnus et ils présentent une très grande variabilité spectrale,
- la construction d'un dictionnaire à partir des spectres de galaxies déjà observées par MUSE ou d'autres instruments, nécessiteraient de rééchantillonner les spectres afin qu'ils correspondent exactement à la grille de pixels des données MUSE,
- de plus la distance des galaxies est inconnue, il faudrait donc dupliquer tous les spectres du dictionnaire pour modéliser le décalage dans le rouge dû à l'éloignement des galaxies.

Il est en revanche possible de chercher à estimer les signatures spectrales des galaxies à partir d'un dictionnaire de composantes simples (composantes continues sinusoïdales, distribution de dirac pour modéliser une raie d'émission, etc). Ce type d'approche, proposé dans le cadre des données MUSE, est présenté dans les paragraphes 1.3.3.1 et 1.3.3.2.

De nombreuses méthodes d'extraction des pôles de mélange et d'estimation des abondances qui se basent sur le modèle de mélange linéaire (1.1) ont été proposées dans la littérature. Ces méthodes peuvent généralement se décomposées en deux grandes étapes :

1. **Extraction des pôles de mélange :** Les hypothèses de somme à un des abondances en chaque pixel et la non négativité des abondances permet de modéliser géométriquement les données dans un simplexe [1] de dimensions N où N représente le nombre de composantes spectrales, *i.e.* dans le cas qui nous intéresse, le nombre de sources à détecter. Les méthodes les plus populaires, NFINDR proposée par Winter [1999], PPI (pour *pixel purity index*) introduite par Boardman [1993], ou encore VCA (pour *vertex component analysis*) présentée dans Nascimento and Bioucas Dias [2005] projettent les données dans un sous-espace de dimension $N-1$ par différentes approches afin de réduire la dimension du problème et pouvoir extraire les pôles de mélange. La plupart des méthodes nécessitent l'hypothèse de présence de pixels purs pour chacun des pôles de mélange, ce qui n'est réaliste dans le cas des données MUSE que pour les galaxies brillantes et spatialement étendues. Les très faibles rapports signal à bruit des galaxies d'intérêts constituent une autre limitation pour ces algorithmes.
2. **Estimation des abondances par inversion du problème :** Une fois les signatures spectrales des pôles de mélanges estimées, il suffit d'inverser le modèle de mélange linéaire (1.1) où la matrice $\mathbf{S}$ contient dans chaque colonne la signature spectrale d'un pôle de mélange extrait à l'étape précédente.

Le principe des méthodes d'estimation des pôles de mélanges et des abondances est difficilement applicables aux données MUSE. En effet, dès l'étape de réduction de la dimension des données, la taille des données MUSE et le nombre de sources observées (et donc de pôles de mélange différents) sont beaucoup trop grands pour les méthodes classiques. Le nombre moyen de sources dans les cubes de données MUSE est de l'ordre de quelques centaines, ce qui donne une dimension bien trop grande pour appliquer les méthodes de types N-FINDR, OSP, VCA, etc. Ces méthodes ont été élaborées pour des données hyperspectrales avec application pour la télédétection (le capteur est alors dirigé vers le sol) contenant au maximum quelques centaines de longueurs d'onde et au maximum quelques dizaines de composantes différentes. De plus en comparaison avec les données MUSE, les données de télédétection présentent très peu de bruit. Il est en revanche possible d'utiliser une stratégie de démélange spectral localement, sur un sous-cube

1. Un simplexe de dimension N est l'analogie du triangle dans un espace de dimension deux étendu à un espace de dimension $N-1$.

ne contenant que quelques galaxies dont les spectres ont été préalablement estimés. Cela permet, notamment lorsque plusieurs galaxies se recouvrent spatialement, d'estimer leurs étendues. Cette approche a été développée dans Shen et al. [2012] pour séparer des galaxies spectralement différentes qui se recouvrent spatialement dans les données MUSE.

1.2.2 Détection de cible

Les auteurs de Manolakis and Shaw [2002], Manolakis et al. [2003] et Manolakis et al. [2014] proposent une étude complète des différentes stratégies de détection de cible dans les images hyperspectrales développées durant les vingt dernières années. Le terme de *détection de cible* (en anglais dans les articles : *target detection*) est utilisé plutôt que le terme de détection de sources car les approches développées dans la littérature hyperspectrale visent souvent à détecter dans un pixel une signature spectrale connue (ou estimée) en utilisant un test d'hypothèse binaire du type :

$$\begin{cases} \mathcal{H}_0 & : \text{ bruit seul} \\ \mathcal{H}_1 & : \text{ source + bruit} \end{cases} \tag{1.2}$$

Ces méthodes sont orientées autour de la classification spectrale pixel par pixel réalisée à l'aide d'un détecteur permettant de rejeter l'hypothèse $\mathcal{H}_0$ lorsque le seuil de détection est dépassé. La forme des détecteurs dépend alors des hypothèses utilisées (présence ou absence de superposition de sources, connaissance de la statistique des pixels de bruit, etc). Parmi les détecteurs recensés dans Manolakis et al. [2003], beaucoup sont issus de l'écriture du test du rapport de vraisemblance, qui selon les hypothèses utilisées dans (1.2), peut alors prendre différentes formes. On peut par exemple citer le détecteur NMF (pour *normalized matched filter*) ou le détecteur ACE (pour *adaptive cosine/coherence estimator*) de Kelly [1986]. D'autres approches pixelliques sont également présentées dans Nasrabadi [2014]. Le principal problème de ces approches est que la détection d'objets basée uniquement sur leurs caractéristiques spectrales conduit à des tests peu puissants. En effet, les caractéristiques spatiales (forme, dimensions, orientations, interactions avec les autres objets, etc.) peuvent également contribuer de manière significative à la détection des objets dans une image hyperspectrale. De plus dans le cadre des données MUSE, nous ne disposons pas du spectre des objets observés, il est donc difficile d'utiliser ces méthodes pour la détection des galaxies. Il serait en revanche possible, une fois les sources détectées et leur spectre estimés, d'utiliser ces méthodes pour affiner la détection de l'étendue spatiale de chacune des sources. Cependant étant donné la dimension des données MUSE (plus de 90000 spectres) et le nombre de galaxies moyen dans un tel cube (quelques centaines), ces traitements doivent être appliqués localement.

1.2.3 Une approche objet

Bien que les méthodes de détection de cible et de démélange spectral soient majoritairement pixelliques, une stratégie de détection d'objets dans les images hyperspectrales est proposée dans les travaux de Valero et al. [2011]. Dans cette approche il ne s'agit pas seulement d'estimer la répartition des différentes signatures spectrales dans chacun des pixels de l'image, mais plutôt de trouver quels sont les ensembles de pixels adjacents (régions) qui correspondent aux caractéristiques spatiales et spectrales du type d'objets recherchés. Un descripteur du type d'objets que l'on cherche à détecter est défini de manière générique par $\mathcal{D} = \{\mathcal{D}_{forme}, \mathcal{D}_{spectral}, \mathcal{D}_{région}\}$. Les pixels de l'image sont tout d'abord classifiés en régions à l'aide d'un arbre de répartition binaire (voir Valero et al. [2010]). Les noeuds de l'arbre sont ensuite parcourus des feuilles jusqu'à la racine, et chaque noeud est analysé à l'aide du descripteur $\mathcal{D}$. Si le noeud ne répond pas favorablement aux critères du descripteur, il est retiré de l'arbre. Après parcours complet de l'arbre de

répartition binaire, les noeuds restant constituent autant d'objets détectés. Les auteurs proposent dans Valero et al. [2011] deux applications : la détection de route et la détection de bâtiment dans une image hyperspectrale. Il est à noter cependant que la détection d'objets développée dans ces applications visent à détecter plusieurs objets du même type, *i.e* avec les mêmes caractéristiques spectrales notamment. Dans le cadre de la détection de galaxies dans les données MUSE, s'il est envisageable d'utiliser les mêmes descripteurs spatiaux pour toutes les sources (un support spatial elliptique), il est plus difficile de modéliser la composante spectrale par un modèle unique. Cette méthode ne pourra donc pas être appliquée directement à la détection de galaxies dans les données MUSE.

1.3 La détection de sources en astrophysique

La détection de sources est un problème classique en astrophysique, de nombreuses méthodes ont été proposées dans la littérature, souvent en association avec un nouvel instrument ou un nouveau jeu de données. Depuis quelques années, de nouveaux instruments produisant des données plus complexes et plus volumineuses voient le jour. Les méthodes classiquement utilisées pour détecter des galaxies sont inadaptées pour traiter ces nouveaux jeux de données (passage de données en deux dimensions à des données de dimensions supérieures, données plus complexes combinant des informations différentes). Dans cette section, nous allons faire un rapide état de l'art des différents types de méthodes que l'on peut trouver dans la littérature et nous verrons que ces méthodes évoluent avec la création de nouveaux instruments.

1.3.1 Les méthodes par seuillage pour les images 2D

Une grande majorité des méthodes de détection de sources dans les images astrophysiques débutent par une étape de seuillage des données qui permet une classification des pixels en deux classes : la classe $\mathcal{C}_0$ qui regroupe les pixels de bruit et la classe $\mathcal{C}_1$ qui rassemble, par opposition, les pixels contenant un signal source. Le seuil permettant cette classification doit être estimé directement à partir des données. La plupart des méthodes appliquent donc une phase d'estimation de la statistique du bruit afin de déterminer un seuil calculé à partir d'une probabilité de fausse alarme ou bien d'un σ-clipping.

1.3.1.1 SExtractor

Avec l'arrivée d'instruments toujours plus performants notamment en terme de taille de champ et de magnitude limite, l'analyse manuelle des images astrophysiques est devenue impossible. Les premiers algorithmes de détection automatique de sources sont apparus vers la fin des années 70. SExtractor (pour Source Extractor) est un logiciel élaboré par Bertin and Arnouts [1996] qui s'inscrit dans la lignée de ces premiers algorithmes. C'est un logiciel dédié à l'extraction automatique des sources astrophysiques dans les images CCD grand champ (jusqu'à 60000×60000 pixels) qui réalise l'analyse en six étapes :

1. **Estimation du bruit de fond du ciel :** Afin de détecter les sources les moins brillantes, il faut estimer le plus précisément possible l'intensité du fond astrophysique en chaque pixel de l'image. Dans cette première étape, le fond est estimé localement. En chaque point de la grille de pixel, l'histogramme local est calculé sur un voisinage puis le mode principal est estimé à l'aide d'une méthode de σ-clipping, *i.e.* l'histogramme est tronqué itérativement à $\pm 3\sigma$ autour de la médiane jusqu'à convergence de la valeur de la médiane. La valeur du fond est ensuite estimée comme la valeur moyenne de cet histogramme tronqué. La carte d'intensité du fond est estimée en appliquant cette méthode à chaque pixel de

l'image, un filtre médian est appliqué afin de lisser le résultat et s'affranchir des éventuelles surestimations du fond du ciel dues à la présence de sources très brillantes.

2. **Seuillage de l'image :** Cette étape permet d'extraire tous les pixels qui se détachent significativement (en accord avec le critère fixé par l'utilisateur passé en paramètre d'entrée de l'algorithme) du fond estimé précédemment. Il faut noter que l'image peut être préalablement filtrée afin d'améliorer la détection : le masque de convolution choisi dépend du type d'objets que l'utilisateur souhaite détecter (sources ponctuelles, sources étendues, etc). A cette étape de la détection, chaque ensemble de pixels connectés (au sens d'un 8-voisinages) est considéré comme une unique source.
3. **Démixage des sources qui se recouvrent :** Cette étape permet de séparer les sources proches spatialement qui ont été détectées comme une unique source. Pour ce faire, chaque ensemble de pixels connectés est seuillé séparément à différents niveau d'intensité (30 niveaux espacés de façon exponentielle entre l'intensité maximale et l'intensité minimale des pixels de l'ensemble considéré). Chaque seuillage est ensuite analysé dans l'ordre décroissant de la valeur des seuils afin de vérifier si l'ensemble de pixels se sépare en plusieurs composantes distinctes.
4. **Filtrage des détections :** Tous les pixels situés au dessus du seuil de détection utilisé lors de la deuxième étape ne correspondent pas forcément à une source. Il faut maintenant filtrer ces détections parasites qui sont souvent situées à proximité d'autres sources détectées. La stratégie mise en oeuvre dans SExtractor est de modéliser la contribution des sources détectées dans le voisinage avec un profil gaussien étendu qui s'ajuste au mieux aux données. Ces contributions sont ensuite soustraites et l'intensité résultante est analysée pour savoir si elle est toujours supérieure au seuil de détection.
5. **Analyse photométrique :** SExtractor produit ensuite une estimation de la magnitude pour une ouverture circulaire et isophote ainsi que la magnitude totale de chaque source détectée.
6. **Séparation des étoiles et des galaxies :** Vient ensuite une étape de classification des sources afin de séparer les étoiles des galaxies. Pour ce faire, un réseau de neurones entrainé sur des images synthétiques est implémenté dans SExtractor. La synthèse des étoiles et des galaxies utilisées pour la phase d'entrainement est décrite dans l'annexe de l'article de Bertin and Arnouts [1996].

SExtractor est un logiciel qui permet de traiter rapidement les images, il est très souvent utilisé dans la communauté astrophysique pour produire des catalogues d'objets de façon automatique (Hogg et al. [2000], Taniguchi et al. [2007]). Il nécessite cependant d'avoir des connaissances en astrophysique pour pouvoir fixer au mieux tous les paramètres de l'algorithme. En plus du grand nombre de paramètres à fixer pour traiter une seule image, il faudrait dans le cas des données hyperspectrales MUSE traiter séparément toutes les images constituant le cube, puis faire une étape de fusion des 3600 catalogues (voir Richard et al. [2015] pour une approche similaire). Cependant, dans le cas des galaxies de type Lyα de faible intensité, la cohérence spectrale n'est pas exploitée par cette approche, et certaines de ces galaxies risquent de ne pas être détectées avec un traitement longueur d'onde par longueur d'onde. Nous verrons qu'il est en revanche possible d'utiliser SExtractor pour extraire des objets de l'image blanche correspondant au cube MUSE.

1.3.1.2 SFIND 2.0 : segmentation de l'image par contrôle du taux de fausses découvertes

Lorsque la distribution des pixels de bruit peut-être modélisée, soit à partir de connaissances physiques, soit à partir d'une estimation (par exemple avec une approche par bootstrapping Efron

and Tibshirani [1994]), le seuil utilisé pour segmenter l'image peut être relié à une probabilité de fausse alarme (se reporter à l'annexe A pour une définition détaillée des notions statistiques évoquées par la suite). Cependant cela ne permet pas de contrôler le taux de fausses découvertes dans la liste de pixels de la classe $\mathcal{C}_1$, *i.e.* la proportion de pixels pour lesquels l'hypothèse de bruit seul à été rejetée à tort parmi tous les pixels de la classe de pixels sources. Dans leurs travaux, Hopkins et al. [2002] proposent une méthode de détection de sources dans les images astrophysiques SFIND 2.0, dont l'étape de seuillage des données est basée sur le contrôle du taux de fausses découvertes, appelé FDR pour *false discovery rate* en anglais, notion introduite par Benjamini and Hochberg [1995] et qui est décrite en annexe dans le paragraphe A.2.2. Le seuillage des images astrophysiques par contrôle du FDR avait été introduite précédemment dans les travaux de Miller et al. [2001]. Les auteurs suggèrent que cette méthode pourraient remplacer la traditionnelle étape de seuillage dans d'autres méthodes de détection telles que SExtractor pour le traitement de grandes images.

La méthode SFIND est composée de plusieurs étapes :

1. **Normalisation des données :** elle est réalisée par ajustement d'une distribution gaussienne sur l'histogramme des pixels de chaque région de l'image. Les régions sont définies par l'utilisateur, il sera donc important d'avoir une bonne connaissance des données afin de choisir de manière judicieuse la taille de ces régions (en prenant garde par exemple d'avoir une répartition suffisamment homogène des pixels, *i.e.* avoir suffisamment de pixels de bruit comparé aux pixels sources). Cette étape permet d'obtenir une image avec des caractéristiques statistiques (moyenne et variance) uniformes. La répartition finale est alors approchée par une gaussienne de moyenne nulle et de variance $\sigma^2 = 1$.
2. **Seuillage par contrôle du FDR :** sous l'hypothèse nulle, *i.e.* l'hypothèse que le pixel ne contienne que du bruit, le pixel est considéré comme une réalisation d'une variable aléatoire distribuée selon une loi gaussienne de moyenne nulle et de variance $\sigma^2 = 1$. Pour chacun des N pixels de l'image, il est alors possible de calculer la p-valeur correspondante (voir annexe A). Ces p-valeurs sont ensuite ordonnées dans l'ordre croissant, et la procédure de seuillage de Benjamini and Hochberg [1995] est alors appliquée. Il faut noter que Hopkins et al. [2002] utilisent une modification du critère de contrôle du FDR proposée par Benjamini and Yekutieli [2001] afin de prendre en compte la corrélation des pixels dues à la réponse impulsionnelle de l'instrument. Ce choix sera discuté dans le chapitre 3 consacré aux tests multiples.
3. **Estimation de la position des sources :** une fois la valeur de seuil calculée, tous les pixels dont la p-valeur est inférieure à ce seuil sont analysés. Les pixels adjacents sont identifiés comme appartenant à une même source, et un algorithme de recherche d'un maximum local dans chaque ensemble de pixels est appliqué. Chaque ensemble de pixels est ensuite modélisé par une gaussienne à deux dimensions dont les paramètres sont ajustés au sens des moindres carrés. Si cette procédure d'ajustement ne converge pas (trop peu de pixels adjacents supérieur au seuil de détection), la source est rejetée.

D'après l'évaluation des performances réalisée dans Hopkins et al. [2002], SFIND 2.0 et SExtractor semblent fournir des performances similaires en terme de sources non détectées et de fausses détections.

1.3.2 Les méthodes développées pour les images en 3D

La détection de sources (étoiles, galaxies) en astrophysique est une problématique de longue date. Nous avons vu dans le paragraphe précédent que la communauté a développé des méthodes performantes pour la détection et l'extraction de sources dans des grands images, comme SExtractor (Bertin and Arnouts [1996]) ou SFIND 2.0 (Hopkins et al. [2002]). La détection de sources

dans des données en trois dimensions est en revanche un problème assez récent, l'apparition d'instruments (MUSE, ASKAP pour *Australian Square Kilometre Array Pathfinder*) produisant des données massives en trois dimensions, deux dimensions spatiales et une dimension spectrale, entraîne une demande forte en méthodes de détection automatique de sources. Nous allons présenter dans ce paragraphe deux méthodes récemment proposées pour le traitement des données de l'instrument ASKAP, mis en service en 2012, dont l'une des missions scientifiques est l'étude de la formation et de l'évolution des galaxies dans l'Univers proche à l'aide de la raie d'émission de l'hydrogène neutre, noté HI, située dans le domaine radio à une longueur d'onde de 21cm.

1.3.2.1 DUCHAMP

L'algorithme DUCHAMP, (Whiting [2012]), est une méthode développée pour détecter des sources dans des données en trois dimensions, dont deux dimensions spatiales et une dimension spectrale (longueur d'onde ou fréquence), donc des données similaires aux observations MUSE. Cette méthode a été élaborée en prévision du grand nombre de données produites par les instruments du Square Kilometre Array (SKA), qui font des observations dans le domaine radio pour la détection des émissions de l'hydrogène neutre, noté HI. Cependant cette méthode est suffisamment générique pour être utilisée sur d'autres données astrophysiques en trois dimensions contenant des sources qui émettent à certaines fréquences. DUCHAMP fournit en sortie une liste de positions des sources détectées dans l'image. L'approche utilisée pour la détection est une analyse pixellique suivie d'une étape de fusion des pixels détectés sans *a priori* sur la nature, la forme ou le profil d'intensité des sources. Les entrées de l'algorithme sont les données et un fichier contenant un grand nombre de paramètres qui permettent d'ajuster les prétraitements des données et les différentes étapes de la détection. Les différents prétraitements (filtrage de la composante continue, reconstruction par ondelettes) sont destinés à réduire le bruit et améliorer la détection des sources les plus faibles.

L'étape majeure du processus de détection est l'étape de seuillage des données, c'est-à-dire la classification des pixels dans la classe $\mathcal{C}_0$ (bruit) et dans la classe $\mathcal{C}_1$ (sources). Le calcul du seuil est important puisque dans la version de la méthode DUCHAMP présentée dans Whiting [2012], le seuil est identique pour toutes les images du cube. Il existe plusieurs façons de calculer ce seuil :

- sous la forme d'un flux minimal,
- comme un rapport signal à bruit avec un seuillage à $n\sigma$, $n \in \mathbb{N}^*$, où σ est l'écart-type du bruit qui peut être estimé par défaut sur l'ensemble des données, ou sur une partie des données qui peut être spécifiée par l'utilisateur dans le fichier de paramètres. Le calcul direct de σ peut être biaisé par les pixels brillants appartenant aux sources. L'utilisation de méthodes plus robustes, comme le calcul de la dispersion absolue médiane : médiane($|x_i -$ médiane$(x_i)|$) permettent d'avoir une estimation plus robuste,
- à l'aide de l'algorithme proposé par Hopkins et al. [2002] (voir paragraphe 1.3.1.2) pour contrôler le taux de fausses découvertes dans la liste de pixels considérés comme appartenant à des objets. Un taux de fausses découvertes q peut être fixé par l'utilisateur dans le fichier de paramètres afin de réaliser la segmentation avec cette approche. Segmenter le cube de données à l'aide de cette méthode nécessite de prendre en compte toutes les précautions citées dans le paragraphe 1.3.1.2.

Une fois la segmentation réalisée, vient l'étape de détection des sources. Dans les images en deux dimensions, l'algorithme de Lutz [1980] est largement utilisé pour détecter des objets à partir de pixels classifiés en une classe $\mathcal{C}_0$ (bruit) et une classe $\mathcal{C}_1$ (sources). L'algorithme parcourt chaque image du cube ligne par ligne en agrégeant les pixels connectés d'une ligne à l'autre au sens d'un 8-voisinage. Cet algorithme est utilisé sur chaque image composant le cube de données. A ce stade l'algorithme produit une liste de sources détectées avec la position spatiale et spectrale

des pixels qui les constituent. Deux méthodes sont implémentées dans DUCHAMP pour réaliser la fusion. La première consiste à considérer que deux sources forment en réalité un seul objet si elles sont adjacentes dans l'une des dimensions du cube, *i.e.* il y a au moins une paire de pixels (un par source) adjacents (au sens d'un 4-voisinage) sur l'une des dimensions du cube. La deuxième approche consiste à considérer que deux sources font partie d'un même objet si elles sont séparées spatialement et spectralement d'un nombre maximum de pixels défini par l'utilisateur (ce nombre peut être différent pour la dimension spatiale et la dimension spectrale).

Une fois les sources détectées à différentes longueurs d'ondes fusionnées, DUCHAMP retourne à l'utilisateur le catalogue définitif des sources et des paramètres les caractérisant. Les objets sont identifiés de trois façons :

— par le pixel contenant la valeur de flux maximale (dans les trois dimensions),
— par la position moyenne (dans les trois dimensions) de l'ensemble de pixels formant la source,
— et par la position moyenne pondérée par le flux de chacun des pixels (dans les trois dimensions).

A ces trois positions viennent s'ajouter le flux maximum et le flux total de chacune des sources, l'extension spatiale, la largeur à mi-hauteur et à 20% de la raie d'émission détectée dans la dimension spectrale.

L'évaluation des performances de DUCHAMP est réalisée dans les travaux de Whiting [2012] et plus particulièrement de Westmeier et al. [2012] et Popping et al. [2012]. Les résultats obtenus sur des cubes synthétiques contenant des sources à différents rapports signal à bruit montrent que malgré des prétraitements efficaces, pour un seuil de détection inférieur à 4σ, DUCHAMP ne peut pas détecter toutes les sources. Il est possible d'utiliser DUCHAMP pour traiter les données MUSE, cependant nous avons observé que la présence de sources ayant plus d'une raie d'émission dans le spectre et une composante continue pouvant osciller fortement, entraine des surdétections aux niveaux de ces sources (les critères de fusion des sources ne sont pas adaptés à d'autres profils spectraux que ceux ne contenant qu'une seule raie d'émission). De plus, étant donnés les faibles rapports signal à bruit des galaxies les plus lointaines, il faut utiliser un seuil de détection inférieur à 4σ, ce qui génère un grand nombre de fausses découvertes (voir la comparaison dans l'article Meillier et al. [2015a]).

1.3.2.2 SoFiA : Source finding application

Très récemment, un nouveau logiciel de détection de sources dans des données en trois dimensions (deux dimensions spatiales et une dimension spectrale) a été mis à disposition par Serra et al. [2015]. Il regroupe différents algorithmes développés pour la recherche de sources émettant de l'hydrogène neutre (HI) dans les données ASKAP, les auteurs affirment cependant que SoFiA peut être utilisé sur des cubes de données provenant d'autres instruments. Il serait donc envisageable de travailler sur les cubes de données MUSE à l'aide de ce logiciel. Nous donnons dans ce paragraphe une courte description des différentes étapes de traitement du logiciel SoFiA :

1. **Prétraitement du cube de données :** puisque tous les algorithmes de détection implémentés dans SoFiA font l'hypothèse que le niveau de bruit est uniforme dans toutes les dimensions du cube (comme l'algorithme DUCHAMP) il faut pouvoir normaliser les données. Pour cela, l'utilisateur peut spécifier un cube de pondération qui correspondrait à l'inverse du niveau de bruit dans chaque pixel du cube de données, ou bien il peut utiliser les méthodes d'estimation du niveau de bruit et de pondération qui sont implémentées dans SoFiA (estimation de l'écart-type, à la moyenne ou à la médiane, ajustement des données par une gaussienne).
2. **Filtrage des données :** avant d'effectuer le seuillage et la détection, une étape de filtrage peut être appliquée afin d'améliorer le rapport signal à bruit.

3. **Seuillage des données :** trois méthodes de seuillage sont implémentées dans SoFiA. La première est un seuillage simple, les valeurs de tous les pixels du cube sont comparées à un seuil spécifié par l'utilisateur et les pixels correspondants sont identifiés comme appartenant à une source. La seconde approche repose sur un algorithme S + C (pour *smooth + clip*) développé par Serra et al. [2012b] qui consiste à filtrer le cube avec différents noyaux de convolution 3D et à seuiller le résultat des différents filtrages (seuil défini comme un rapport signal à bruit). Les pixels détectés à l'aide d'au moins un filtre sont considérés comme appartenant à une source. La troisième méthode consiste à tester la validité de chaque spectre sous l'hypothèse nulle, c'est-à-dire que le pixel considéré ne contient que du bruit.
4. **Fusion des pixels en sources :** Comme dans DUCHAMP l'algorithme de Lutz [1980] est utilisé pour fusionner les pixels adjacents et former les sources en trois dimensions.
5. **Rejet des fausses détections :** Cette étape peut être menée de deux façons, soit en spécifiant un nombre minimal de pixels adjacents (afin de ne pas détecter de sources moins étendues que ne le permet la résolution de l'instrument), soit en estimant la validité des sources sur l'hypothèse que les pics de bruit sont autant à valeurs positives qu'à valeurs négatives si la distribution du bruit est bien symétrique. Cette seconde méthode est détaillée dans Serra et al. [2012a].
6. **Optimisation du masque représentant chaque source :** à l'étape de seuillage, le masque correspondant au cube de données prend la valeur 1 au niveau des pixels supérieur au seuil de détection et la valeur 0 partout ailleurs. Cette étape d'optimisation permet d'élargir ou de corriger le masque au niveau de chaque source en se basant sur un critère de flux total de la source.
7. **Description paramétrique des sources :** Les sources sont ensuite modélisées par leur position (barycentre du flux, position du pic de flux), le flux total, la taille et l'occupation de la source dans les trois dimensions du cube ainsi que d'autres paramètres pouvant être utile à l'analyse des sources.
8. **Résultats de l'algorithme :** en plus du fichier contenant tous les paramètres décrivant les sources, le logiciel fournit entre autre une estimation du spectre, les sous-cubes contenant chaque source et les images correspondant aux différentes projections des sources (image moyenne, champ de vitesse).

SoFiA est un logiciel de détection de sources très flexible puisqu'à chaque étape de la chaine de traitement, l'utilisateur à le choix entre différents algorithmes. L'application aux données MUSE semble possible, SoFiA permettrait notamment d'exploiter les informations du cube de variance. Notons que lors de l'évaluation des performances dans Serra et al. [2015], les auteurs suggèrent de travailler sur des cubes de données de dimensions raisonnables (inférieures à $360 \times 360 \times 1464$ pixels) pour des questions de puissance de calcul et d'occupation mémoire. Or un cube MUSE est deux fois plus volumineux que la taille maximale des cubes utilisés dans Serra et al. [2015] pour évaluer les performances. Il faudra donc disposer d'importants moyens de calcul et de ressources mémoire suffisantes pour traiter un cube MUSE avec le logiciel SoFiA.

1.3.3 La détection de source appliquées aux données MUSE

Nous présentons dans ce paragraphe un rapide état de l'art des méthodes proposées par des équipes du consortium MUSE ou affiliée au consortium via des projets communs. Ces approches ont toutes un fil conducteur différent, l'idée étant non pas de converger vers une unique méthode de détection, mais au contraire de proposer à la communauté astrophysique un ensemble d'outils permettant d'aboutir à un catalogue de sources détectées en croisant les résultats obtenus.

1.3.3.1 Décomposition des spectres sur un dictionnaire, Bourguignon et al. [2011]

La première approche proposée dans la littérature est celle de Bourguignon et al. [2011] où les auteurs s'intéressent au débruitage, à la déconvolution et la détection de sources dans les données MUSE. Les auteurs considèrent le cube de données MUSE comme une collection de spectres qu'ils traitent tout d'abord séparément. Dans leur modèle, un spectre $\boldsymbol{s}$ peut se décomposer comme la somme d'une composante continue $\boldsymbol{s^c}$ et de quelques raies d'émission $\boldsymbol{s^\ell}$:

$$\boldsymbol{s} = \boldsymbol{s^c} + \boldsymbol{s^\ell}$$

La composante continue est décomposée sur une base de sinusoïde pour modéliser les variations lentes et continues du spectre, tandis que les raies d'émissions sont représentées par des pics modélisés à l'aide de distribution de Dirac. Ces deux composantes peuvent se représenter de manière parcimonieuse dans différents espaces de représentation :

- la transformée en cosinus discret de $\boldsymbol{s^c}$ est parcimonieuse. En reprenant les notations du papier de Bourguignon et al. [2011], si $\boldsymbol{W_{DCT}} \in \mathbb{R}^{N\times N}$ est la matrice de transformation en cosinus discret et N le nombre de longueurs d'onde dans le spectre, alors $\boldsymbol{s^c} = \boldsymbol{W_{DCT}}^T \boldsymbol{x^c}$ avec $\boldsymbol{x^c}$ un vecteur dont seuls quelques coefficients sont significatifs.
- la composante modélisant les raies d'émission est parcimonieuse dans la base canonique associée à $\mathbb{R}^N$: $\boldsymbol{s^\ell} = \boldsymbol{x^\ell}$, avec $\boldsymbol{x^\ell}$ un vecteur de coefficients nuls presque partout.

L'estimation du spectre $\boldsymbol{s} = [\boldsymbol{W_{DCT}}^T, I_N][\boldsymbol{x^c}, \boldsymbol{x^\ell}]$ via une pénalisation ℓ_1 permet de débruiter en partie les données. Une fois les spectres débruités, vient une étape d'aggrégation des spectres pour former des sources distinctes. Deux spectres appartenant à des pixels connectés sont dits semblables lorsqu'ils présentent une composante non nulle commune (*i.e.* présence d'une raie d'émission à la même longueur d'onde dans les deux spectres). La segmentation a donc lieu longueur d'onde par longueur d'onde :

- pour chaque longueur d'onde λ_n, $n \in \{1, \dots, N\}$, les pixels connectés (au sens d'un 8-voisinage) qui présentent une composante $\boldsymbol{x^\ell}_{\lambda_n}$ non nulle sont associés. A ce stade, la liste de composantes est redondante dans le sens où une galaxie peut présenter plusieurs raies d'émission et donc être représentée par plusieurs composantes (2D) situées à différentes longueurs d'onde,
- afin d'éliminer les composantes redondantes, si deux composantes partagent une grande proportion de leur extension spatiale alors elles sont fusionnées en une seule composante (3D),

Le nombre final de composantes correspond donc au nombre de sources détectées par cette approche. Une dernière étape d'estimation du spectre global de la source est réalisée à partir des spectres des pixels composant la source. Si cette méthode fournit des résultats satisfaisants, elle présente un coût calculatoire important qui dépend de la dimension des données et du nombre de sources à détecter.

1.3.3.2 Rapport de vraisemblance généralisé sous contrainte, Paris et al. [2013b]

Les travaux de Paris et al. [2013b] et Mary et al. [2014] reposent également sur l'idée que la décomposition d'un spectre à l'aide d'un dictionnaire de composantes spectrales simples est parcimonieuse. Par rapport aux travaux de Bourguignon et al. [2011], le dictionnaire est enrichi avec des modèles de raies d'émission pour modéliser la raie Lyα. Les auteurs proposent deux versions d'un test statistique basé sur le rapport de vraisemblance généralisé sous contrainte de parcimonie. Dans la première version du test, les spectres formant le cube de données MUSE sont traités indépendamment les uns des autres, seule la composante spectrale de la réponse impulsionnelle de l'instrument est prise en compte. Or la réponse impulsionnelle de l'instrument possède une composante spatiale qui dilue l'information dans un voisinage de taille $N_x \times N_y$ (voir la modélisation dans le paragraphe 1.5.1). Dans la seconde version du test, la corrélation

spatiale induite par la réponse impulsionnelle de l'instrument est donc introduite dans le modèle. Cette seconde version rend le test plus puissant (voir les travaux de thèse de Paris [2013]), nous nous intéresserons donc directement à ce test. Alors que dans le premier modèle, seul le spectre $\boldsymbol{s} \in \mathbb{R}^{\Lambda}$ était considéré, dans la deuxième version un sous-cube de taille $N_x \times N_Y \times \Lambda$ centré sur le spectre $\boldsymbol{s}$ à la position spatiale (x, y) est considéré.

Nous allons nous intéresser à cette méthode en détail car elle est très proche méthodologiquement de ce qui sera présenté dans le chapitre 3. En reprenant les notations introduites dans le papier Paris et al. [2013b], $\boldsymbol{s}_V(x, y)$, vecteur de taille $N_x N_Y \Lambda \times 1$, est la version vectorisée de ce sous-cube où les $N_x \times N_Y$ spectres sont stockés les uns en dessous des autres. Sous les deux hypothèses du test, on a donc :

$$\begin{cases} \mathcal{H}_0 & : & \boldsymbol{s}_V(x,y) = \boldsymbol{\epsilon}_V & \text{(bruit seul)} \\ \mathcal{H}_1 & : & \boldsymbol{s}_V(x,y) = \mathbf{F}(\mathrm{x},\mathrm{y})\mathbf{H}\mathbf{R}\boldsymbol{\alpha} + \boldsymbol{\epsilon}_{\mathrm{V}} & \text{(source + bruit)} \end{cases} \tag{1.3}$$

où $\boldsymbol{\epsilon}_V \in \mathbb{R}^{N_x N_Y \Lambda}$ est un bruit gaussien $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_V(x, y))$ avec $\boldsymbol{\Sigma}_V(x, y)$ la matrice de covariance de taille $N_x N_Y \Lambda \times N_x N_Y \Lambda$ construite à partir des informations de variance fournies avec le cube de données MUSE. La composition par la composante spatiale de la réponse impulsionnelle de l'instrument est modélisée par la matrice $\mathbf{F}(\mathrm{x}, \mathrm{y}) \in \mathbb{R}^{\mathrm{N_x N_Y \Lambda \times \Lambda}}$ (voir le calcul dans Paris et al. [2013b]) et la composition par la composante spectrale est modélisée par la matrice $\mathbf{H} \in \mathbb{R}^{\Lambda \times \Lambda}$. La matrice $\mathbf{R} \in \mathbb{R}^{L \times \Lambda}$ représente le dictionnaire des L composantes spectrales simples sur lequel le spectre est décomposé, et le vecteur $\boldsymbol{\alpha} \in \mathbb{R}^{L}$ est le vecteur des coefficients à estimer. Dans le cas de la détection de galaxies lointaines dans les données MUSE, le spectre de ces galaxies est principalement composée d'une raie unique d'émission :

$$\|\boldsymbol{\alpha}\|_0 = 1, \tag{1.4}$$

ce qui se traduit par une décomposition du spectre en une seule composante simple. Après normalisation des données, le test (1.3) se reformule par :

$$\begin{cases} \mathcal{H}_0 & : & \boldsymbol{z}_V(x,y) = \boldsymbol{\Sigma}_V^{-\frac{1}{2}} \boldsymbol{s}_V(x,y) = \boldsymbol{w} & \text{(bruit seul)} \\ \mathcal{H}_1 & : & \boldsymbol{z}_V(x,y) = \boldsymbol{\Sigma}_V^{-\frac{1}{2}} \boldsymbol{s}_V(x,y) = \mathbf{D}\boldsymbol{\theta} + \boldsymbol{w} & \text{(source + bruit)} \end{cases} \tag{1.5}$$

où $\mathbf{D}$ est un dictionnaire équivalent dont les colonnes sont normalisées et $\boldsymbol{\theta}$ est le vecteur de coefficients $\boldsymbol{\alpha}$ pondéré par la matrice diagonale contenant la norme des colonnes de $\mathbf{D}$. La contrainte (1.4) se reformule donc par :

$$\|\boldsymbol{\theta}\|_0 = 1, \tag{1.6}$$

ce qui permet d'écrire sous l'hypothèse $\mathcal{H}_1$:

$$\boldsymbol{z}_V(x,y) = \mathbf{d}_{\mathrm{j}} \boldsymbol{\theta}_{\mathrm{j}} + \boldsymbol{w} \quad , \quad \|\boldsymbol{\theta}\|_0 = 1 \tag{1.7}$$

où $\boldsymbol{\theta}_j$ est le seul élément non nul du vecteur $\boldsymbol{\theta}$ et $\mathbf{d}_{\mathrm{j}}$ la composante du dictionnaire $\mathbf{D}$ associée. Le rapport de vraisemblance généralisé s'écrit finalement sous la forme :

$$\mathrm{GLR}_{\mathrm{1s}}^{(\mathrm{3D})} = \frac{\max\limits_{\boldsymbol{\theta}} \mathrm{p}(\boldsymbol{z}_{\mathrm{V}}(\mathrm{x},\mathrm{y})|\mathbf{D}, \boldsymbol{\theta})}{\mathrm{p}(\boldsymbol{z}_{\mathrm{V}}(\mathrm{x},\mathrm{y})|\boldsymbol{\theta} = \mathbf{0})} = \frac{\max\limits_{\mathrm{j}, \boldsymbol{\theta}_{\mathrm{j}}} \mathrm{p}(\boldsymbol{z}_{\mathrm{V}}(\mathrm{x},\mathrm{y})|\mathbf{d}_{\mathrm{j}}, \boldsymbol{\theta}_{\mathrm{j}})}{\mathrm{p}(\boldsymbol{z}_{\mathrm{V}}(\mathrm{x},\mathrm{y})|\boldsymbol{\theta} = \mathbf{0})} \tag{1.8}$$

et donc le test associé s'écrit :

$$T_{\mathrm{GLR}_{\mathrm{1s}}^{(\mathrm{3D})}}(\boldsymbol{z}_V(x,y)) = \max_j \left| \mathbf{d}_{\mathrm{j}}{}^{\mathrm{T}} \boldsymbol{z}_{\mathrm{V}}(\mathrm{x},\mathrm{y}) \right| \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\lessgtr}} \xi \tag{1.9}$$

où ξ est un seuil calculé selon la procédure présentée dans Paris et al. [2011] et qui dépend de la probabilité de fausse alarme fixée pour ce test. Le test (1.9), appelé par la suite max-test

en références aux travaux de Arias-Castro et al. [2011] qui ont introduit ce test, est appliqué à chaque spectre du cube de données MUSE. Le résultat final est une carte binaire de détection où chaque position (x, y) détectée est supposée appartenir à une source avec une probabilité de fausse alarme [2] correspondant au seuil ξ utilisé. Pour être exhaustif, le dictionnaire de composantes spectrales simples doit contenir les $\Lambda \simeq 3600$ répliques translatées de chaque composante spectrale puisque la distance des galaxies étant inconnue, le décalage dans le rouge de leur spectre est à priori inconnu. Même en se restreignant à la détection de galaxies lointaines dont le spectre contient une unique raie Lyα, étant donné la variabilité de la forme de cette raie, le catalogue reste de dimension trop grande pour pouvoir réaliser l'estimation de chaque spectre du cube (d'après les auteurs, il faudrait plus de 6 mois pour traiter tout le cube avec un catalogue de 10000 profils spectraux qu'il faudrait translater autour de chacune des 3600 longueurs d'onde formant le cube). L'idée développée dans Paris et al. [2013b] repose sur la construction d'un dictionnaire de taille réduite à partir d'un catalogue de profil de raies Lyα fournies par les astrophysiciens afin d'améliorer la puissance du max-test (1.9) et réduire la probabilité de fausse alarme liées à un trop grand nombre de composantes possibles pour représenter un spectre. Finalement, pour la détection de raie Lyα, les auteurs proposent différents catalogues contenant une composante translatée pour les Λ longueurs d'onde lorsqu'il est réduit par SVD, $7 \times \Lambda$ composantes lorsqu'il est réduit par K-SVD et $1 \times \Lambda$ lorsqu'il est réduit par minimax (voir Suleiman et al. [2013]). Finalement le catalogue utilisé contient des atomes 3D (sous forme vectorisée), *i.e.* le profil spectral composé par la réponse impulsionnelle en trois dimensions de l'instrument MUSE.

Si cette méthode permet de construire une carte (dans la dimension spatiale du cube de données MUSE) des pixels susceptibles d'appartenir à une source, des critères de fusion des pixels détectés et de séparation des sources lorsqu'il y a recouvrement spatial de plusieurs galaxies restent à mettre en oeuvre. Des travaux de thèse (C. Clastres) s'inscrivant dans la continuité des travaux de Paris [2013] sont maintenant en cours afin de réaliser l'agglomération des pixels détectés, c'est-à-dire de passer d'une détection pixellique à une représentation par sources.

Dans le chapitre 3 de ce manuscrit, nous verrons que le test appliqué à la sortie d'un outil très classique du traitement du signal, le filtrage adapté, nous conduit à une formulation similaire au test (1.9) proposé par Paris et al. [2013b].

1.3.3.3 Méthodes de détections développées au sein du consortium MUSE

D'autres méthodes sont en cours de développement au sein du consortium MUSE : la méthode CubEX présentée par Cantalupo [2014], et la méthode LSDcat développée par Herenz [2014]). Ces deux méthodes sont pour le moment internes au consortium [3] et n'ont pas encore fait l'objet de publications.

1.4 Les données

Afin de mesurer les performances des différentes méthodes de détection de sources dans des données hyperspectrales comme celles de MUSE, les astrophysiciens du consortium ont fournit des simulations numériques sensées ressembler aux cubes de données réelles fournies par MUSE. L'une de ces simulations a servi durant la première partie de la thèse avant de basculer sur des données réelles acquises une fois MUSE installé au foyer de l'UT4 du VLT.

2. Se reporter au paragraphe A.1.2.2 dans l'annexe A pour la définition du lien entre probabilité de fausse alarme et seuil de décision.

3. Ces méthodes ont été présentées au consortium lors de la busy week MUSE en novembre 2014.

1.4.1 Les données synthétiques : le DryRun

Il existe au sein du consortium un modèle numérique de l'instrument MUSE. Dans ce simulateur l'ensemble de la chaine d'acquisition (via le télescope et l'instrument MUSE) est modélisée ainsi que la chaine de traitement des données en sortie des 24 modules composant l'instrument. Grâce à ce modèle numérique présenté dans Jarno et al. [2008] et grâce aux modèles physiques des galaxies, quelques cubes de données synthétiques ont pu être ainsi produits afin de tester et valider les algorithmes de traitement de données élaborés avant la mise en marche de l'instrument. Ces cubes de données synthétiques sont appelés *dry runs* ; l'élaboration et l'étude de l'un de ces *dry runs* sont détaillées dans Husser [2012]. Le cube synthétique sur lequel nous allons travailler est un cube de taille 100×100 pixels pour les dimensions spatiales avec un échantillonnage spatiale de 0.2 arcseconde et de 3600 longueurs d'onde réparties sur l'intervalle $[480\text{nm}, 930\text{nm}]$ avec un échantillonnage de 0.125nm. Ce cube contient 16 galaxies et 2 étoiles de tailles, de formes et d'intensités différentes représentant les différents types de sources que l'on devrait rencontrer dans les données réelles. Dans la suite de ce manuscrit nous appellerons ce cube DryRun. La figure 1.6 est l'image blanche obtenue en sommant les images obtenues pour les différentes longueurs d'onde avant l'ajout du bruit. Nous donnons dans le tableau 1.1 un descriptif des objets présents dans ce cube. Afin de tester les performances des algorithmes de détection [4], ce cube se devait d'être le plus représentatif possible des observations réalisées par l'instrument MUSE. Ainsi ce cube contient une configuration de galaxies illustrant différents cas de détection délicats :

- deux galaxies, ID#14 et ID#15, qui se recouvrent partiellement,
- idem pour ID#11 et ID#12,
- une étoile, ID#2, qui recouvre partiellement une galaxie, ID#16,
- des galaxies, ID#1 et ID#5, qui ont un profil d'intensité assez irrégulier.

Le bruit additif est parfaitement gaussien, de moyenne nulle sur toutes les longueurs d'onde. En revanche la variance de ce bruit varie fortement avec la longueur d'onde comme le montre la figure 1.7. La détection d'une galaxie avec une raie d'émission de faible intensité aux longueurs d'onde où le bruit a une forte puissance sera impossible sans traitement des données visant à améliorer le rapport signal à bruit, en exploitant des informations spectrales et spatiales sur la source à détecter.

Etant donné la grande variabilité de la puissance du bruit dans ce cube de données, il serait incohérent de considérer un RSB moyen pour caractériser la détectabilité des différentes sources présentes dans le DryRun. En effet pour une galaxie dont le spectre est constitué d'une unique raie d'émission, même si l'intensité dans les longueurs d'onde correspondantes est très forte, le moyennage sur les 3600 longueurs d'onde mènerait à un RSB très faible. De plus, deux galaxies distinctes présentant chacune une raie d'émission de même intensité et de même largeur, mais située à deux longueurs d'onde différentes, ne seront pas caractérisées par le même niveau de rapport signal à bruit compte tenu de la grande variabilité de la puissance du bruit. Nous préférons donc définir le RSB d'une source s_i centrée en la position spatiale (x_i, y_i) comme le rapport signal à bruit à la longueur d'onde correspondant à la valeur maximale du RSB calculé en chaque longueur d'onde tel qu'il est défini dans l'équation (1.10) donnée ci après.

$$\text{RSB}(s_i) = \max_{\lambda} 10\log\left(\frac{s_i(x_i, y_i, \lambda)^2}{\sigma_\lambda^2}\right), \tag{1.10}$$

où $s_i(x_i, y_i, \lambda)$ est la valeur du signal à la position (x_i, y_i, λ), et σ_λ^2 est estimé par la variance du cube de bruit (données bruitées - vérité terrain) dans ce cas synthétique. Dans le cas réel, nous serons obligés d'utiliser soit le cube de variance Σ_{MUSE} fourni avec les données s'il est fiable,

4. Le DryRun a été utilisé dans d'autres travaux sur la détection de sources dans les données MUSE, notamment dans les travaux de Paris et al. [2013a] et Mary et al. [2014].

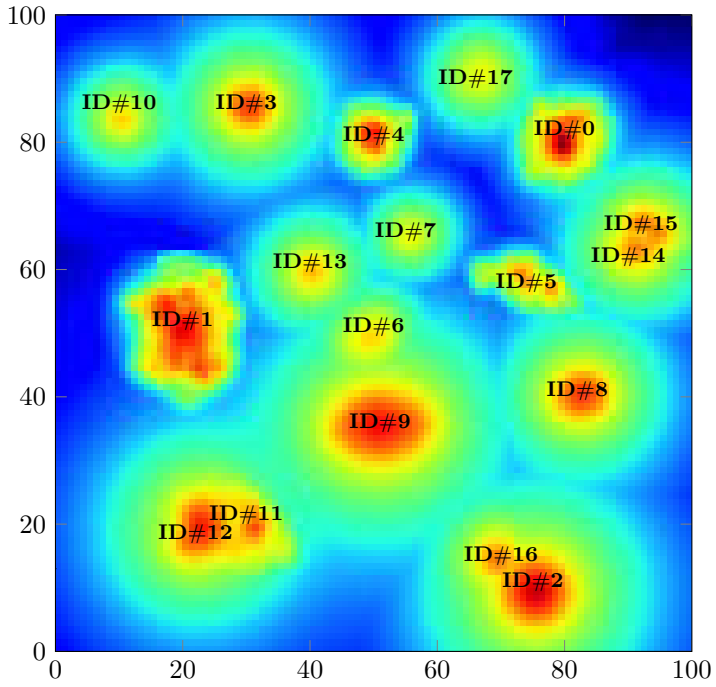


FIGURE 1.6 – Image blanche déduite du cube DryRun sans bruit (obtenue en sommant les images du cube selon l'axe des longueurs d'ondes). Les identifiants ID#NN permettent d'identifier par la suite les objets.

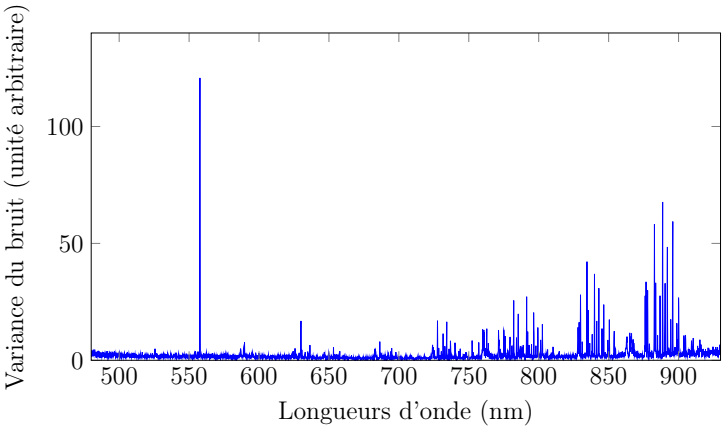


FIGURE 1.7 – Variance empirique du bruit additif gaussien en fonction de la longueur d'onde.

soit une estimation robuste de σ^2_λ, par exemple par σ-clipping, sur l'image correspondant à la longueur d'onde λ.

Les figures 1.8, 1.9, 1.10 et 1.11 correspondent à l'évolution du rapport signal à bruit en fonction de la longueur d'onde pour différents types de sources du DryRun.

ID	(x,y)	CARACTERISTIQUES SPECTRALES	RSB
0	(80,80)	continu + raie d'émission OII (908.3nm)	30.7 dB
1	(20,50)	continu + raie d'émission OII (906.8nm)	28.8 dB
2	(75,10)	continu (étoile)	34.2 dB
3	(30,85)	continu (étoile)	21.5 dB
4	(50,80)	continu + raie d'émission OII (759.1nm)	20.7 dB
5	(74,57)	continu + raie d'émission OII (870.7nm)	3.3 dB
6	(50,50)	Lyα (734.2nm)	-1.5 dB
7	(55,65)	Lyα (870.3nm)	-6.9 dB
8	(82,40)	Lyα (772.8nm)	17.7 dB
9	(50,36)	Lyα (531.8nm)	24.7 dB
10	(10,85)	Lyα (833.6nm)	-7.26 dB
11	(30,20)	continu	6.8 dB
12	(22,20)	Lyα (531.8nm)	22.3 dB
13	(40,60)	Lyα (508.7nm)	0.5 dB
14	(90,62)	Lyα (480.3nm)	4.2 dB
15	(92,65)	Lyα (845.5nm)	-1.9 dB
16	(70,14)	Lyα (517.8nm)	7.7 dB
17	(66,90)	Lyα (585.7nm)	-5.5 dB

TABLEAU 1.1 – Caractéristiques des 18 sources présentes dans le cube DryRun. Les positions (x, y) sont données en pixels et la position dans le spectre des raies d'émissions est indiquée dans la colonne "caractéristiques spectrales". Le RSB indiqué pour chaque objet correspond à la valeur maximale obtenue sur le spectre du centre de chacun des objets, voir la définition donnée par l'équation (1.10). Le bruit additif est celui des données bruitées par les astrophysiciens, il est supposé gaussien, spatialement stationnaire, et indépendant pixel à pixel dans les dimensions spectrale et spatiale.

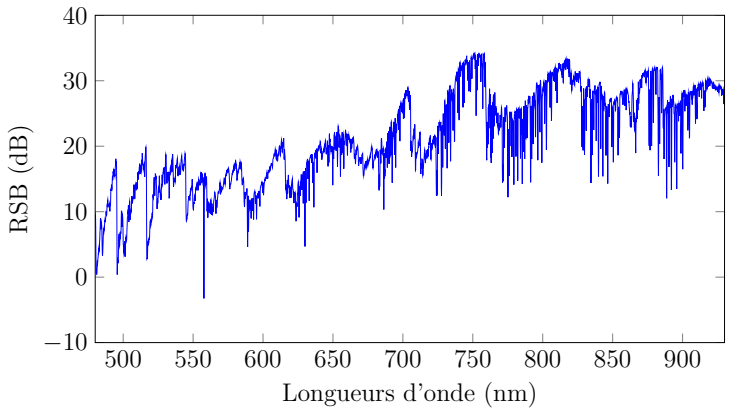


FIGURE 1.8 – Evolution du RSB local en longueur d'onde pour l'étoile ID#2 dont le spectre est composée d'une composante continue de forte intensité.

Dans le cas des galaxies de type émetteur Lyα (figures 1.9 et 1.10) nous constatons que si l'intensité au niveau de la raie d'émission permet d'atteindre des RSB compris entre -7.26dB pour la galaxie la moins brillante (source #10) et 24.7dB pour la galaxie la plus brillante (source #9), l'absence de signal sur le reste du spectre conduit à des RSB très faibles (entre -90 et -10dB). Il

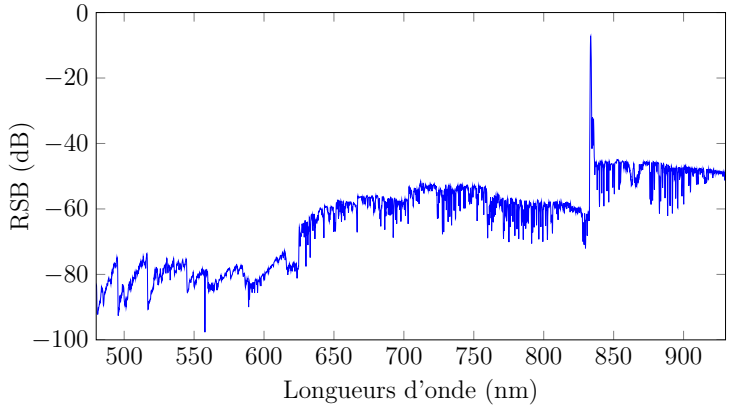


Figure 1.9 – Evolution du RSB local en longueur d'onde pour la galaxie lointaine ID#10, on observe un RSB maximal à la longueur d'onde $\lambda = 833.6$nm correspondant à la raie d'émission Lyα de faible intensité.

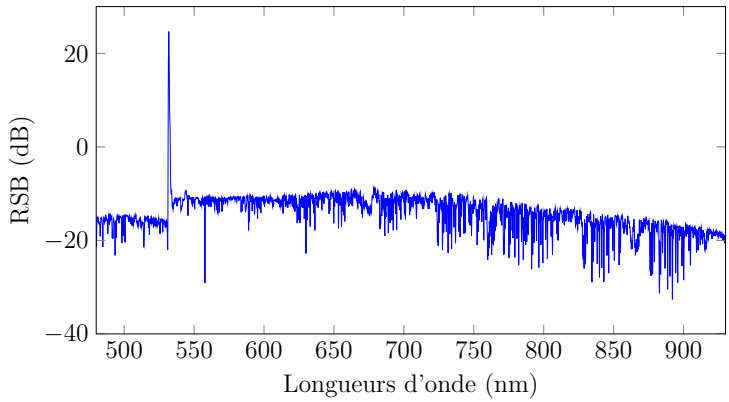


Figure 1.10 – Evolution du RSB local en longueur d'onde pour la galaxie ID#9, on observe un RSB maximal à la longueur d'onde $\lambda = 531.8$nm correspondant à la raie d'émission Lyα très brillante.

faut également prendre en compte le fait que ces RSB sont donnés pour le centre des sources et que la décroissance spatiale d'intensité est souvent très rapide lorsque l'on s'écarte du centre de la galaxie. La détection des galaxies comme la source #10 ne sera donc pas aisée avec un RSB maximal de -7.26dB et une extension spatiale et spectrale très faible. Nous montrerons dans le chapitre 3 comment améliorer la détectabilité de ce type de sources en prétraitant les données.

1.4.2 Les données réelles : le Hubble Deep Field South

Afin d'obtenir un cube de données réelles qui puisse servir à l'évaluation des performances de l'instrument MUSE et des méthodes de détection de galaxies dans les champs profonds, il faut pouvoir observer avec MUSE une portion du ciel bien connue des astrophysiciens et dont les sources observées sont relativement bien caractérisées par la communauté astrophysique. Le choix s'est porté sur une portion du champ Hubble Deep Field South que nous allons décrire dans le paragraphe suivant.

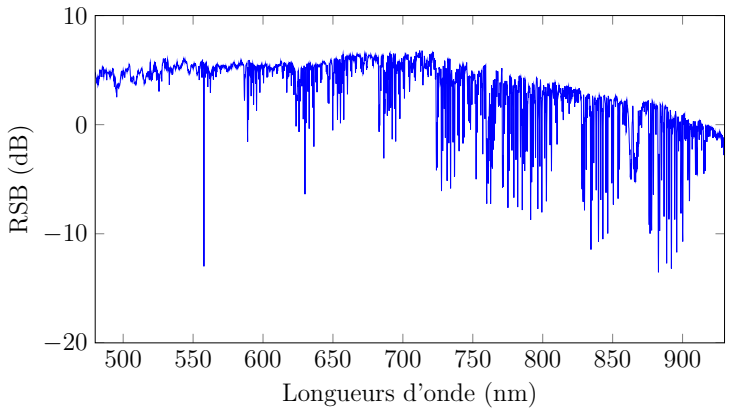


FIGURE 1.11 – Evolution du RSB local en longueur d'onde pour la galaxie ID#11 qui n'a qu'une composante spectrale continue, d'intensité quasiment constante sur toute la gamme de longueurs d'onde considérée.

1.4.2.1 Le champ observé par Hubble

Le télescope spatial Hubble a observé deux champs profonds : le Hubble Deep Field North (HDFN) en 1995 et le Hubble Deep Field South (HDFS) en 1998 (voir la description détaillée dans les travaux de Ferguson et al. [2000]). Réaliser un champ profond consiste à collecter la lumière pendant un très long temps de pose (plusieurs dizaines d'heures) afin d'enregistrer suffisamment de photons provenant des sources de plus faible intensité et pouvoir ainsi les détecter sur l'image finale. Nous nous intéresserons ici au HDFS puisque MUSE est installé dans l'hémisphère sud. L'image du HDFS datant de 1998 est la combinaison de plusieurs centaines d'images issues de différentes poses avec des filtres sélectifs en longueur d'onde différents (ici 300nm, 400nm, 606nm et 814nm) réparties sur 10 jours entre septembre et octobre 1998. La résolution de l'image finale est de 0,0398 arcseconde, elle contient plusieurs milliers d'objets (galaxies, étoiles, et un quasar). L'image du HDFS est présentée sur la figure 1.12.

1.4.2.2 Le champ observé par MUSE et la construction des données

Durant la dernière phase de mise en service de l'instrument MUSE à Paranal, la portion du champ HDFS (1×1 arcmin^2) illustrée sur la figure 1.12 a été observée par MUSE (Bacon et al. [2015]). Il aura fallu au total 54 poses de 30 minutes chacune, répartie sur huit nuits, soient 27h de temps d'intégration, pour réaliser le cube HDFS hyperspectral avec MUSE.

Nous donnons ici un rapide résumé de l'ensemble de la chaîne de traitement qui permet de passer des 54 cubes des poses individuelles au cube final, les étapes sont détaillées précisément dans Bacon et al. [2015] et dans Weilbacher et al. [2012] :

1. Compilations des mesures réalisées sur les poses de calibration (biais, arcs, champ plat).
2. Correction des défauts à l'aide des données de calibration et transformation des coordonnées du détecteur en coordonnées du ciel.
3. Astrométrie : calibration du flux sur toutes les poses, correction des petits décalages entre les poses introduits par l'oscillation du dérotateur de champ, etc.
4. Elimination des *systématiques* (offsets entre les slices) en ramenant toutes les slices autour de la même valeur médiane.
5. Interpolation par drizzling en trois dimensions (algorithme adapté pour les trois dimensions de MUSE à partir des travaux de Fruchter et al. [2009]) qui permet d'aligner les 54

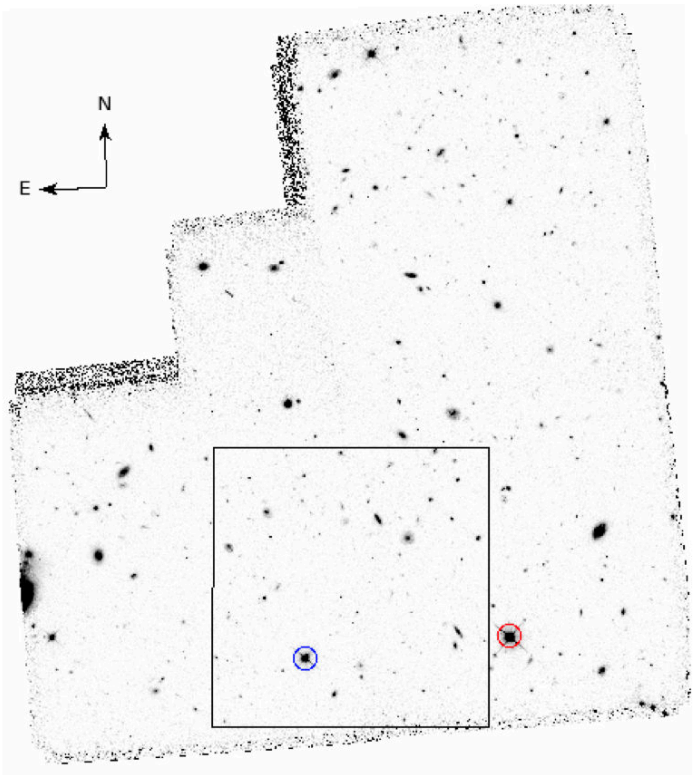


FIGURE 1.12 – Le Hubble Deep Field South réalisé par le télescope spatial Hubble et la portion de ce champ observé par MUSE en juillet 2014 modélisée par le carré noir.

poses individuelles sur la même grille de pixels que le cube final. Chaque position (x, y, λ) de la grille de pixels subit ensuite une opération de σ-clipping (à 5σ) sur les 54 poses afin de rejeter les poses dont le pixel considéré est contaminé par un rayon cosmique dans le calcul final du pixel considéré. Ainsi chaque pixel du cube sera une moyenne d'un nombre N de pixels correspondant aux N poses individuelles conservées par le σ-clipping pour le pixel considéré. Ce nombre N varie de pixel à pixel.

6. Soustraction de la contribution du ciel (voir le spectre d'émission sur la figure 1.13) sur chacune des 54 poses individuelles à l'aide du logiciel ZAP développé par Soto et al (papier en préparation) dans le cadre du consortium MUSE.
7. Les 54 poses sont ensuite fusionnées pour former le cube final. En parallèle du cube final, est produit également un cube de variance Σ_{MUSE}, où la variance de chaque pixel (x, y, λ) est estimée à partir des N valeurs retenues par le σ-clipping :

$$\Sigma_{MUSE}(x, y, \lambda) = \frac{1}{N-1} \sum_{i=1}^{N} \left(\mathbf{Y}_i(x, y, \lambda) - \overline{\mathbf{Y}_i(x, y, \lambda)} \right)^2 \tag{1.11}$$

où $\mathbf{Y}_i$ est le cube de données de la i^{me} pose individuelle et $\overline{\mathbf{Y}_i(x, y, \lambda)}$ est la valeur du cube de données final obtenu par moyennage des N pixels de coordonnées (x, y, λ) des N poses retenues par le σ-clipping.

Si d'un point de vue astrophysique tous ces traitements visent à supprimer des effets physiques indésirables, d'un point de vue traitement du signal, certaines opérations de la chaine de réduction des données induisent des effets non négligeables sur les données. L'opération de drizzling en trois

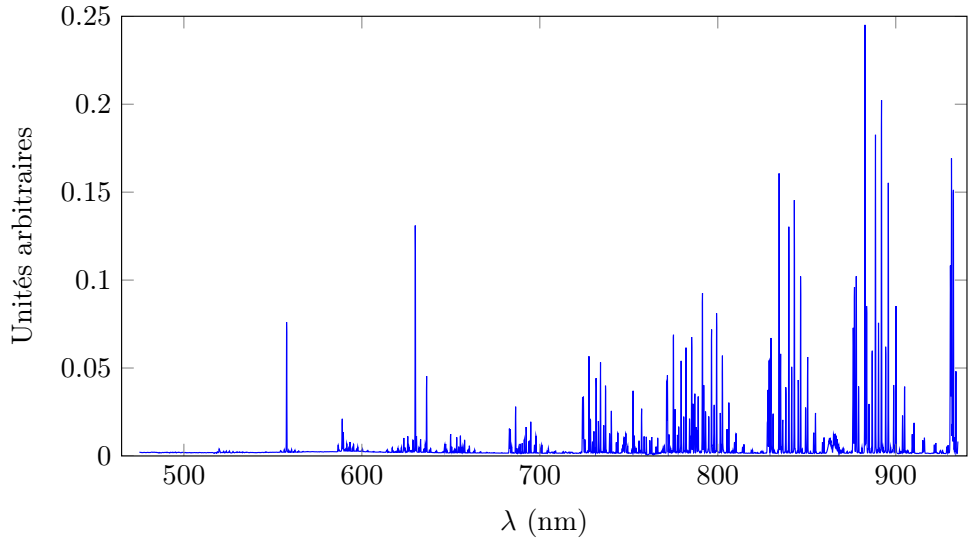


FIGURE 1.13 – Spectre du ciel dans la gamme de longueur d'onde de l'instrument MUSE.

dimensions induit par exemple une certaine corrélation (spatiale et spectrale) des pixels avec leurs proches voisins ; il faudra prendre en compte cet effet lors de la formulation d'hypothèses et de modèles notamment pour les pixels de bruit (*i.e.* ne contenant pas de contribution d'une source). La soustraction du ciel, imparfaite du fait des erreurs de modélisation de la réponse impulsionnelle spectrale de l'instrument MUSE, a également des répercussions sur le niveau de bruit à certaines longueurs d'onde. La figure 1.14 correspond au spectre moyen d'une zone de bruit du cube HDFS. La contribution du ciel dans ce spectre est réduite lorsque l'on compare au spectre illustré sur la figure 1.13 (les deux spectres sont représentés sur la même échelle d'amplitude normalisée, en unités arbitraires). Cette soustraction est nécessaire, notamment à la détection de raies Lyα situées dans la partie rouge du spectre ($\lambda \geqslant 700$ nm) cependant l'opération entraine la présence de nombreuses valeurs négatives et positives de forte amplitude dans le spectre aux abords des raies d'émission du ciel. Cette étape de la chaine de traitement des données est toujours en cours d'évolution afin d'améliorer la qualité de la soustraction de ciel. Pour le moment ces résidus de la soustraction du fond de ciel sont absorbés dans ce que nous modéliserons comme le bruit de fond (voir paragraphe 1.5.2).

1.4.2.3 Catalogue d'objets

Le HDFS est l'un des champs profonds les plus étudiés dans la littérature. Les astrophysiciens du consortium MUSE ont inspecté les données afin d'identifier dans le cube HDFS de MUSE les galaxies précédemment observées dans l'image HDFS de Hubble et compléter les informations relatives à ces sources en mesurant un redshift précis. Finalement c'est un catalogue de 189 sources (181 galaxies et 8 étoiles) qui a été établi et dont la répartition est présentée dans le tableau 1.2

Ce catalogue ne contient pas toutes les sources détectables présentes dans le cube HDFS de MUSE, notamment des galaxies de type émetteurs Lyα de faible intensité. En effet, lorsque la raie d'émission Lyα n'est pas parfaitement identifiable et mesurable, la source est écartée de ce catalogue d'objets. Ce dernier doit être fiable afin de mesurer les performances des algorithmes de détection. Nous disposons également du catalogue des sources détectées sur l'image HDFS par Casertano et al. [2000]. Il contient un plus grand nombre de sources (586) du fait de la meilleure résolution spatiale du télescope Hubble et de toutes les sources dont le spectre contient

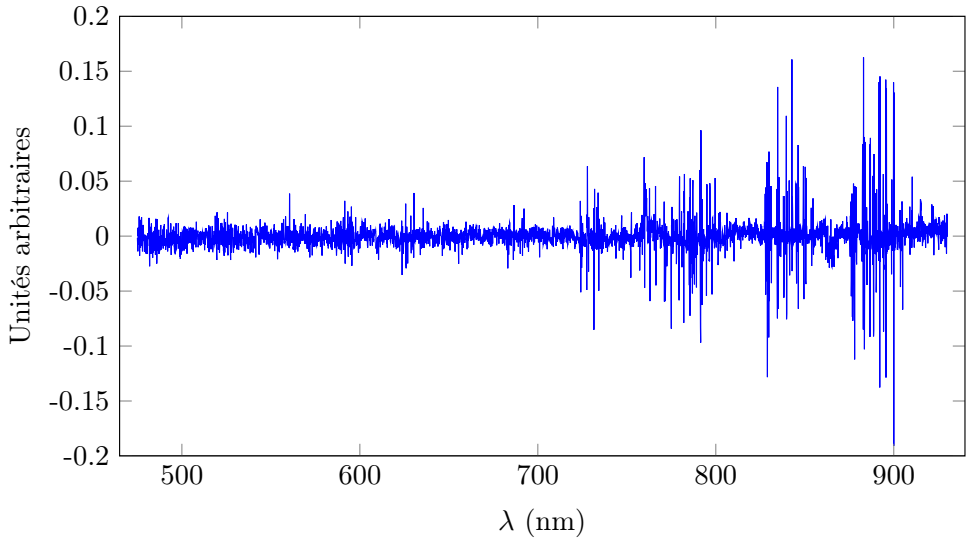


FIGURE 1.14 – Spectre moyen d'un sous-cube du champ HDFS ne contenant aucun objet après les opérations de soustraction du ciel. De nombreux résidus, positifs et négatifs, subsistent du fait du sous échantillonnage et de la variabilité de la réponse impulsionnelle spectrale de l'instrument MUSE qui entraine des erreurs de modélisation des raies du ciel dans le cube de données.

Catégorie	**Nombre**	**Redshift**
étoiles	8	0
galaxies proches	7	0.12 - 0.28
émetteurs OII	61	0.29 - 1.48
galaxies avec raies d'absorption	10	0.83 - 3.90
galaxies à noyau actif (AGN)	2	1.28
émetteurs CIII	12	1.57-2.67
émetteurs Lyα	89	2.96-6.28

TABLEAU 1.2 – Répartition des sources du catalogue d'objets observés et mesurés dans le cube HDFS.

une composante continue sans raie d'émission identifiable et qui ne sont donc pas répertoriées dans le catalogue des 189 objets avec redshifts mesurés.

1.5 Modélisation des données et de la configuration de galaxies

Les données MUSE sont complexes, afin de réaliser la détection de galaxies dans ces champs de données, il nous faut établir un modèle de galaxies et une modélisation du bruit simple mais suffisamment robuste. Il faut également tenir compte de la réponse impulsionnelle de l'instrument MUSE qui peut être estimée, voire mesurée sur les observations et sur les données de calibration.

1.5.1 Modélisation de la réponse impulsionnelle de l'instrument

La PSF (fonction d'étalement du point, ou en anglais *point spread function*) de l'instrument MUSE est définie dans les trois dimensions du cube de données : deux dimensions spatiales et une dimension spectrale. C'est une fonction complexe qui intègre les effets de l'atmosphère, de

l'instrument lui-même et du télescope. Nous nous appuyons sur les travaux de Carfantan [2014], Villeneuve et al. [2011], Serre et al. [2010] et notamment sur les hypothèses simplificatrices qu'ils ont établis pour modéliser la PSF de MUSE.

La PSF de MUSE n'est pas invariante spatialement et spectralement. Il est donc nécessaire de la définir pour chaque position 3D du cube hyperspectral. Notons $h_{z_p,z_q,\mu}(p,q,\lambda)$ la PSF centrée sur le pixel de coordonnées (z_p, z_q, μ) où (z_p, z_q) sont les coordonnées spatiales et μ est la coordonnée spectrale. C'est une fonction de trois variables : p et q pour les dimensions spatiales [5] et λ pour la dimension spectrale. Du fait de sa variation spatiale et spectrale, l'influence de la PSF sur un signal quelconque $x(p,q,\lambda)$ s'exprime donc comme une composition suivant l'opérateur de Fredholm, noté ici $\otimes$:

$$\boldsymbol{y}(p,q,\lambda) = (x \otimes h)(p,q,\lambda) = \sum_{z_p}\sum_{z_q}\sum_{\mu} x(z_p,z_q,\mu)h_{z_p,z_q,\mu}(p,q,\lambda) \tag{1.12}$$

Il y a autant d'expressions différentes de la PSF $h_{z_p,z_q,\mu}(p,q,\lambda)$ que de positions (p,q,λ) dans le cube de données (soit plus de 300 millions de pixels). Nous comprenons donc la nécessité d'introduire des hypothèses simplificatrices pour la modélisation de cette PSF.

L'hypothèse la plus importante concernant la PSF de l'instrument a été introduite, étudiée et justifiée par les travaux de Carfantan [2014], Villeneuve et al. [2011], Serre et al. [2010], elle peut être résumée par la propriété suivante :

Séparabilité : *La PSF 3D de l'instrument MUSE peut se séparer comme le produit d'une composante spatiale et d'une composante spectrale.*

Cette hypothèse a été largement adoptée par le consortium MUSE, elle repose sur des considérations physiques : l'atmosphère est principalement responsable de l'étalement spatial tandis que l'étalement spectral est induit par l'instrument. La PSF spatiale est appelée FSF pour *field spread fonction* et la PSF spectrale est appelée LSF pour *line spread function*. La forme la plus générale que nous pouvons donner à la FSF centrée sur le point (z_p, z_q, μ) est la fonction $F_{z_p,z_q,\mu} : (p,q) \mapsto F_{z_p,z_q,\mu}(p,q)$. De même pour la LSF exprimée comme la fonction $L_{z_p,z_q,\mu} : \lambda \mapsto L_{z_p,z_q,\mu}(\lambda)$. Nous pouvons définir une hypothèse simplificatrice pour la LSF et la FSF :

Invariance dans le champ d'observation : *La FSF et la LSF sont invariantes dans le champ d'observation spatial de l'instrument, nous pouvons donc écrire :* $F_{z_p,z_q,\mu}(p,q) = F_\mu(p - z_p, q - z_q)$ *et* $L_{z_p,z_q,\mu}(\lambda) = L_\mu(\lambda)$.

Il s'agit d'une hypothèse forte, en particulier pour la LSF qui varie dans le champ (voir les travaux de Carfantan [2014] et la thèse Villeneuve [2012]), mais en l'absence d'optique adaptative, cette hypothèse semble réaliste et elle permet la simplification du problème. L'invariance par translation dans le champ permet de réduire l'expression (1.12) de la composition par la PSF de l'instrument à :

$$\boldsymbol{y}(p,q,\lambda) = (x \otimes h)(p,q,\lambda) = \sum_{z_p}\sum_{z_q}\sum_{\mu} L_\mu(\lambda)F_\mu(p - z_p, q - z_q)x(z_p,z_q,\mu) \tag{1.13}$$

Ainsi pour caractériser complétement la PSF de l'instrument il suffit de connaitre les Λ expressions de la LSF ($\Lambda \simeq 3600$) et les Λ expressions de la FSF. Nous remarquons également que l'on peut réécrire l'équation (1.13) comme une convolution classique entre la FSF et le signal puis une composition par la LSF (qui du fait de sa variabilité spectrale ne peut s'écrire comme une convolution dans le domaine spectral) :

$$\boldsymbol{y}(p,q,\lambda) = (x \otimes h)(p,q,\lambda) = \sum_{\mu} L_\mu(\lambda)\Big\{\Big(F_\mu * x(\cdot,\cdot,\mu)\Big)(p,q)\Big\}$$

5. Pour des raisons pratiques les notations (p,q) seront utilisées à la place des notations classiques (x,y) pour désigner les coordonnées spatiales afin de ne pas confondre avec $\boldsymbol{y}$ qui désigne le vecteur contenant les données et x qui désigne la réponse spatiale normalisée d'une source dans la suite de ce document.

1.5.1.1 Modélisation de la FSF

En l'absence d'optique adaptative la FSF est modélisée par une fonction de Moffat (Moffat [1969]), c'est un modèle régulièrement utilisé pour modéliser la PSF spatiale en astrophysique, voir par exemple Trujillo et al. [2001]. Elle a pour expression :

$$F_\lambda(p,q) = F_0\Big(1+\frac{p^2+q^2}{\alpha_\lambda^2}\Big)^{-\beta_\lambda}$$

où F_0 est l'intensité au centre de la fonction, *i.e.* pour $(p,q)=(0,0)$ et α_λ et β_λ sont les coefficients de la fonction Moffat. L'expression de la FSF de MUSE varie en fonction de la longueur d'onde λ ce qui explique la dépendance en λ des coefficients α_λ et β_λ. La largeur à mi-hauteur, abrégée FWHM, pour *Full-Width at Half Maximum*, s'exprime de la façon suivante :

$$\mathrm{FWHM}_\lambda = 2\alpha_\lambda\sqrt{2^{\frac{1}{\beta_\lambda}}-1},$$

elle est notamment utile pour définir la résolution spatiale de MUSE.

Les paramètres de la FSF sont estimés grâce aux travaux de Villeneuve et al. [2011]. Dans le module de code correspondant, la FSF est estimée à partir de l'étoile la plus brillante dans le champ d'observation. Les paramètres de la fonction de Moffat sont déterminés de manière à optimiser la modélisation du profil d'intensité de cette étoile par une fonction de Moffat bi-variée (définie sur un sous espace de $\mathbb{R}^2$). Le paramètre β_λ ne varie quasiment pas avec la longueur d'onde alors que la FWHM montre une décroissance en λ.

D'après Bacon et al. [2015], les mesures réalisées à partir de l'étoile située dans le champ du cube HDFS ont permit d'établir que la FWHM de la FSF décroit de façon linéaire en fonction de la longueur d'onde selon l'équation suivante :

$$FWHM_\lambda = -4.5454\times10^{-4}\lambda + 0.9781 \tag{1.14}$$

où λ est la longueur d'onde exprimée en nanomètres. Le paramètre β est quant à lui estimé à une valeur constante de $\beta = 2.6$. Le paramètre α_λ est déduit directement à partir des valeurs $FWHM_\lambda$ et β, nous pouvons ainsi construire le gabarit de la FSF sur toutes les longueurs d'onde. Pour le traitement du cube HDFS, nous considérons une FSF de 21×21 pixels de large afin d'avoir une décroissance continue du profil de la FSF vers zéro et limiter les effets de bord lors de la convolution avec ce profil. L'évolution de la FSF de MUSE est représentée en deux dimensions sur la figure 1.15 et en coupe sur la figure 1.16 dans le cas du cube HDFS.

1.5.1.2 Modélisation de la LSF

La LSF est mesurée directement sur les données de calibration de l'instrument. Cette LSF varie en fonction de la longueur d'onde, cependant, elle est estimée en moyenne avec une largeur à mi-hauteur de 2.1 ± 0.2 échantillons spectraux (Bacon et al. [2015]). La LSF est sous-échantillonnée ce qui explique l'incertitude sur sa mesure. Cependant cette incertitude ne devrait pas avoir de répercussion majeure sur la détection de galaxies et sur l'estimation des spectres des objets puisque ceux-ci présentent des caractéristiques spectrales plus larges que la LSF. Le modèle utilisé pour la LSF de l'instrument MUSE est une gaussienne échantillonnée dont la variance varie légèrement avec la longueur d'onde, on donne une représentation de cette LSF pour différentes longueurs d'onde dans la figure 1.17.

1.5.2 Modélisation du bruit

Les données MUSE en champ profond sont des données extrêmement bruitées au regard des galaxies que nous cherchons à détecter. Le bruit présent dans les données MUSE est le résultat

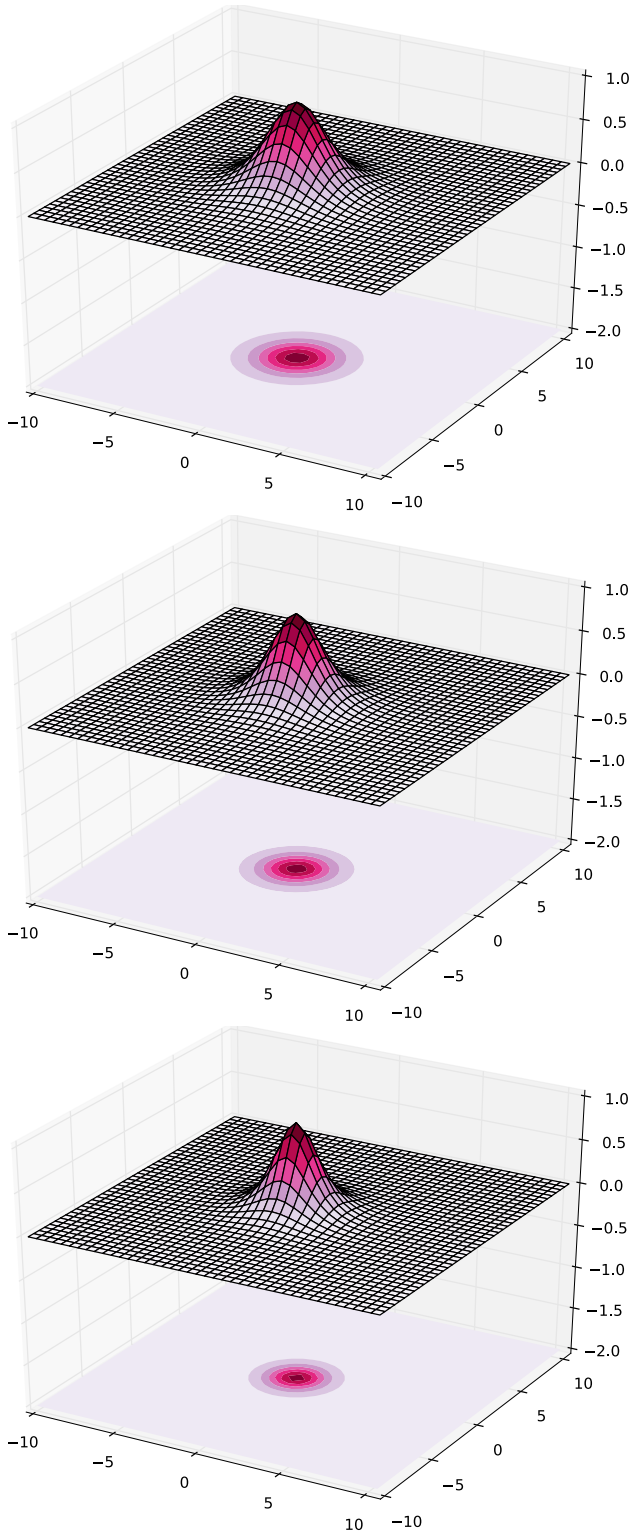


FIGURE 1.15 – Modélisation de la FSF (non normalisée) de MUSE pour le cube HDFS à différentes longueurs d'onde. Haut : $\lambda = 480nm$ ($\text{FWHM}_{480} = 0.76$ arcsec), centre : $\lambda = 700nm$ ($\text{FWHM}_{700} = 0.66$ arcsec) et bas = $\lambda = 930nm$ ($\text{FWHM}_{930} = 0.55$ arcsec).

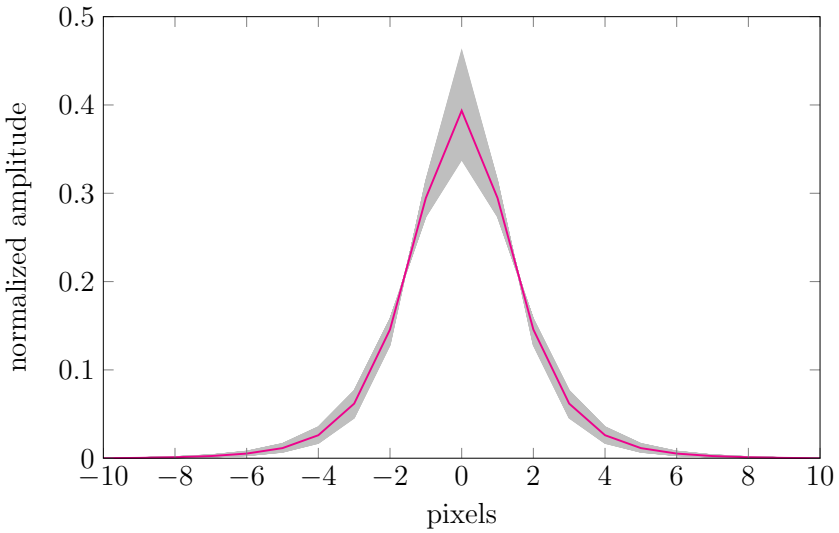


FIGURE 1.16 – Vue en coupe de la modélisation de la FSF (normalisée selon la norme ℓ_2) de MUSE pour le cube HDFS à différentes longueurs d'onde. En gris sont représentées les fonctions Moffat sur une coupe transversale qui modélise la FSF pour les différentes longueurs d'onde et en magenta est représentée la FSF moyenne.

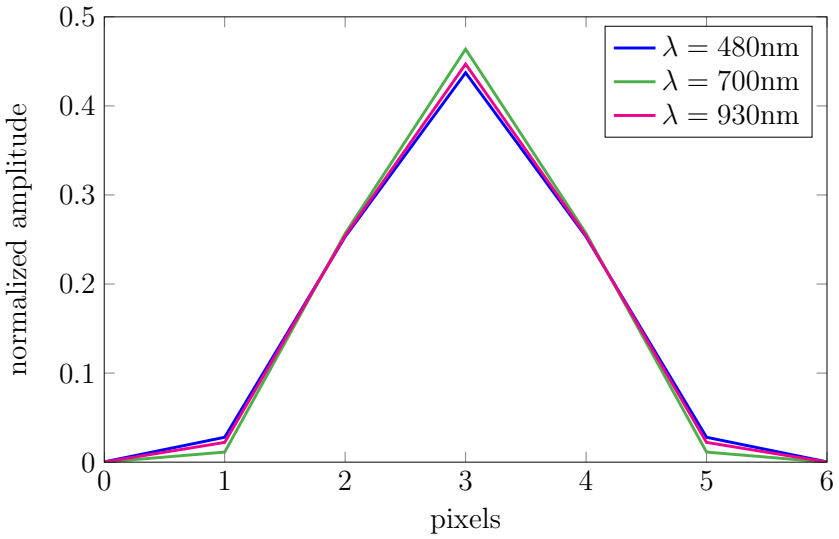


FIGURE 1.17 – Modèle de la LSF de MUSE pour différentes longueurs d'onde

de plusieurs sources de bruit qui se mélangent et détériorent la contribution des galaxies dans le signal. Le bruit est principalement dû aux émissions parasites de l'atmosphère, on retrouve notamment un fort niveau de bruit au niveau des longueurs d'onde correspondant aux raies d'émission du ciel (voir le spectre caractéristique du ciel sur la figure 1.13). Nous avons vu dans le paragraphe 1.4.2 qu'une étape de soustraction de ciel avait lieu dans la chaine de traitement qui conduit à l'élaboration du cube de données final. Cette soustraction génère des résidus sur le spectre du ciel (voir figure 1.14). Ces résidus sont une contribution supplémentaire au bruit présent dans les données MUSE. L'instrument lui-même est également un générateur de bruit : température des capteurs, incertitudes sur les mesures, etc, sont résumés dans un bruit de mesure. Et enfin la nature poissonnienne du flux de photons qui impactent le détecteur engendre une variation spatiale et spectrale du niveau de bruit qui sera plus élevé là où le flux lumineux sera plus important (notamment au niveau des étoiles et des galaxies très brillantes).

La modélisation de toutes ces contributions de bruit adoptée dans le consortium MUSE repose sur un modèle de bruit additif gaussien dont la moyenne est constante spatialement, mais dont la variance varie dans les trois dimensions du cube. Un cube de variance estimée en chaque pixel (x, y, λ) selon le processus décrit dans le paragraphe 1.4.2 est fourni avec les données MUSE.

Sur les données réelles, nous avons vu que la construction d'un cube de données nécessite la combinaison de plusieurs poses individuelles grâce à un algorithme de drizzling. Cette étape engendre une corrélation au niveau des pixels d'un voisinage en trois dimensions. Cependant la corrélation induite n'est pas parfaitement caractérisée, il est difficile de suivre la corrélation de pixels à priori indépendants de leurs voisins dans les cubes des poses individuelles, après l'étape de drizzling. Il faudrait propager un cube de covariance par pixel, ce qui étant donné les dimensions d'un cube de données MUSE est impossible. Dans le modèle de mélange linéaire présenté dans le chapitre 2 de ce manuscrit, nous allons considérer un modèle de bruit simplifié. Nous avons choisi de décomposer les données en une somme de contributions des galaxies observées et d'un bruit qui englobent toutes les sources de bruit citées précédemment ainsi que les erreurs de modélisation des galaxies, notamment des plus brillantes. Ce bruit sera considéré comme identiquement et indépendamment distribué selon une loi gaussienne de moyenne m_λ et de variance σ_λ^2 constantes dans le plan d'observation à la longueur d'onde λ. Considérer la moyenne et la variance constantes à une longueur d'onde λ donnée revient à négliger le bruit de photons liés aux galaxies, notamment au niveau des plus brillantes où nous supposons que les erreurs de modélisation du profil d'intensité des galaxies sont plus fortes que les erreurs engendrées par l'hypothèse de stationnarité dans le champ spatial. De plus puisque le ciel a été soustrait au niveau des poses individuelles, dans les zones de bruit du cube final, il ne doit subsister aucun signal qui aurait pu être étalé spectralement par la LSF de l'instrument. Nous supposons également que la corrélation spectrale des pixels de bruit peut être négligée, c'est une hypothèse forte puisque cette corrélation est non nulle même sur les zones du cube ne contenant aucune source, mais elle est nécessaire à la résolution du problème de détection puisque nous ne disposons d'aucune information sur la corrélation induite par le drizzling et que l'estimation de cette structure de corrélation, *a priori* non isotrope et non stationnaire, s'avère délicate. Finalement à une longueur d'onde λ donnée, le bruit additif est distribué selon :

$$\boldsymbol{\epsilon}_{Bg,\lambda} \sim \mathcal{N}(m_\lambda, \sigma_\lambda^2 \mathbf{I}_M) \tag{1.15}$$

où $\boldsymbol{\epsilon}_{Bg,\lambda}$ est un vecteur de taille $M \times 1$ correspondant à l'image vectorisée du bruit additif à la longueur d'onde λ et M est le nombre de pixels dans une image du cube. Et dans la dimension spectrale, le vecteur de bruit $\boldsymbol{\epsilon}_{Bg}(p, q) \in \mathbb{R}^\Lambda$, situé à la position (p, q), est distribué selon :

$$\boldsymbol{\epsilon}_{Bg}(p, q) \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{\Sigma}) \tag{1.16}$$

où $\boldsymbol{m} = [m_1, \cdots, m_\Lambda]$ est le vecteur de taille $\Lambda \times 1$ contenant l'évolution de la moyenne du bruit

avec la longueur d'onde et la matrice $\mathbf{\Sigma} \in \mathbb{R}^{\Lambda \times \Lambda}$ est diagonale dont les coefficients de la diagonale sont les variances du bruit aux longueurs d'onde correspondantes : $\text{diag}(\mathbf{\Sigma}) = [\sigma_1^2, \cdots, \sigma_\Lambda^2]$.

1.5.3 Modélisation spatiale des galaxies

Dans la littérature, les galaxies sont couramment modélisées à l'aide de composantes elliptiques pour décrire le support spatial des galaxies (Bertin and Arnouts [1996], Simard et al. [2002], Peng et al. [2002]). Concernant la modélisation de la décroissance d'intensité du centre de la galaxie vers les bords, les modèles font souvent appel aux profils Sersic introduit par Sersic [1963] comme une généralisation du profil introduit par De Vaucouleurs [1959].

Dans l'algorithme SExtractor développé par Bertin and Arnouts [1996], il n'existe pas de modèle pour le profil d'intensité des galaxies, les sources sont modélisées comme une collection de pixels adjacents. Cependant l'algorithme peut renvoyer, en plus de cet ensemble de pixels, une estimation elliptique du support constitué par cet ensemble de pixels.

L'approche GIM2D (pour *Galaxy IMage 2D*) développée par Simard et al. [2002] permet de modéliser les galaxies lointaines de faible rapport signal à bruit en une somme de profils d'intensité. Une galaxie peut être modélisée par une ou deux composantes elliptiques avec les profils suivants :

- un profil Sersic lorsque la galaxie est modélisée par une seule composante elliptique
- un profil de De Vaucouleur pour le noyau et un profil exponentiel pour le disque (deux composantes elliptiques)
- un profil Sersic pour le noyau et un profil exponentiel pour le disque (deux composantes elliptiques)

GIM2D prend en entrée une imagette à deux dimensions contenant une galaxie (la segmentation d'une image contenant plusieurs galaxies est réalisée au préalable avec le logiciel SExtractor de Bertin and Arnouts [1996]) et fournit en sortie une image de la galaxie extraite et un catalogue de paramètres caractérisant cette source. L'estimation des paramètres modélisant la galaxie est réalisée par échantillonnage de la fonction de vraisemblance par méthode de Monte Carlo avec un algorithme de Metropolis (Metropolis et al. [1953]).

L'algorithme Galfit proposé par Peng et al. [2002] s'inscrit dans la lignée des algorithmes de type GIM2D qui permettent de modéliser les galaxies dans des images (en deux dimensions). Cependant, la modélisation est plus fine puisque Galfit permet la modélisation de galaxies complexes en autorisant le mélange d'un nombre quelconque de fonctions paramétriques (Sersic, Moffat, gaussienne, exponentielle, etc). Il peut, de plus, modéliser simultanément plusieurs galaxies présentes sur une même image. L'estimation se fait par un algorithme d'optimisation au sens des moindres carrés non linéaire qui implémente la méthode de Levenberg-Marquardt. Il faut noter que cette méthode nécessite une image de variance associée à l'image à analyser. Cette image de variance est l'équivalent du cube de variance fourni avec les données MUSE, elle peut cependant être estimé à partir des caractéristiques de l'image (gain, bruit de mesure) si elle n'est pas estimée d'une autre façon. Galfit prend également en compte la PSF de l'instrument, dont il faut fournir l'image comme paramètre d'entrée de l'algorithme. Les profils d'intensité sont décrits dans un repère elliptique généralisé (voir Peng et al. [2010] pour une description de ce repère) et le profil peut ensuite être déformé afin de représenter par exemple la morphologie des galaxies spirales. Cette modélisation complexe est particulièrement bien adaptée aux images de galaxies proches et donc bien résolues. Dans les champs de données profonds de MUSE, la majorité des galaxies sont lointaines et donc faiblement résolues (voire non résolues), il n'est dont pas utile d'utiliser un modèle aussi complexe pour représenter les galaxies observées. Il ne faut cependant pas exclure l'utilisation de cette méthode pour modéliser quelques galaxies proches qui pourraient se trouver dans le champ de données et fournir le modèle obtenu en guise d'initialisation à l'algorithme de détection des galaxies lointaines que nous proposons dans ces travaux de thèse.

1.5.4 La modélisation des galaxies dans les données MUSE

La stratégie adoptée dans cette thèse consiste à modéliser de manière parcimonieuse la distribution spatiale des galaxies et à utiliser des modèles simples pour décrire une galaxie. Dans ce contexte, les approches par processus objets semblent toutes désignées pour modéliser la configuration de galaxies observées.

Un processus ponctuel marqué est un processus aléatoire dont les réalisations sont des configurations aléatoires d'objets. Chaque objet est représenté par un point (sa position dans l'image) auquel sont associées des marques qui correspondent à ses propriétés géométriques, spectrales, d'intensité, etc. L'avantage de ce type de processus est de fournir une représentation intrinsèquement parcimonieuse de la configuration d'objets recherchée, *i.e.* de la configuration de galaxies dans le cas des données MUSE. Les processus ponctuels marqués permettent de se rapprocher du modèle physique en délaissant le modèle numérique qui revient à considérer le cube de données comme un tableau de pixels en trois dimensions. La modélisation par processus ponctuel marqué permet de lier la nature continue des observations avec la représentation pixellique du cube de données (tableau de pixels en trois dimensions qui représente l'observation échantillonnée et quantifiée). Elle permet de plus de s'affranchir des problèmes de complexité calculatoire inhérents aux approches pixelliques (du fait de la dimension du cube de données), dans le cadre de la modélisation par processus ponctuel marqué, la dimension du problème ne dépendra plus que du nombre d'objets effectivement présent dans la configuration et du nombre de paramètres associés à chaque objet.

Nous avons choisi d'estimer les paramètres décrivant la configuration d'objets, mais aussi la distribution du bruit à l'aide d'une approche bayésienne. Dans ce cas, les paramètres inconnus sont considérés comme des variables aléatoires dont il faut spécifier la distribution : la distribution *a priori*. A l'aide du théorème de Bayes, nous pouvons ensuite exprimer la densité conditionnelle, appelée *distribution a posteriori*, des paramètres sachant les données observées. Ceci permet de faire un compromis entre l'information sur les paramètres du modèle disponible *a priori* (avant de prendre en compte les observations) et les informations apportées par des observations sur ces paramètres. L'utilisation d'une approche bayésienne entraîne un coût de modélisation et de calcul [6] plus élevé que les méthodes de détection par seuillage (SExtractor, DUCHAMP, SOFIA) mais cela se justifie par l'aptitude de ce type de méthode à prendre en compte les modèles de distribution des paramètres inconnus.

Le cadre théorique dans lequel s'inscrivent les processus ponctuels marqués est détaillé dans l'annexe B et la description des galaxies inspirée des modèles utilisés dans GIM2D est détaillée dans le chapitre 2.

6. La formulation bayésienne d'un problème d'estimation demande souvent de calculer des intégrales qui n'ont pas d'expression analytique, et qui nécessitent d'être approchées par simulation.

1.6 Bilan

Les données MUSE

Ce sont des données hyperspectrales massives avec plus de 3600 bandes spectrales, et un champ d'observation de plus de 300×300 pixels pour les dimensions spatiales. Elles contiennent de nombreuses sources avec une grande variabilité spectrale :

- des sources à spectre continu (étoiles, galaxies proches),
- des sources avec une unique raie d'émission (galaxies lointaines),
- des sources dont le spectre contient un composante continue et des raies d'émission (galaxies).

Deux cubes de données seront étudiés dans ce manuscrit :

- le **DryRun** : jeu de données synthétiques contenant 18 sources différentes avec différents rapports signal à bruit et différentes structures spatiales (sources étendues et sources quasi-ponctuelles),
- le **HDFS** : jeu de données réelles acquis par MUSE en 2014. Ce champ a été précédemment observé par le télescope spatial Hubble et un catalogue de sources est disponible pour réaliser des comparaisons avec les résultats de détection.

Le cahier des charges de la méthode de détection

Les méthodes de détection de sources existantes ne permettent pas de répondre au cahier des charges, à savoir, réaliser la détection des galaxies par une approche objet, en trois dimensions, afin de s'affranchir des limitations des approches pixelliques. La stratégie adoptée repose sur :

- une **modélisation simple** des galaxies à l'aide d'un **processus ponctuel marqué**,
- l'estimation des paramètres des objets dans un **cadre bayésien** non informatif.

Chapitre 2

La méthode de détection

Sommaire

Ce chapitre est dédié à la méthode de détection de galaxies par processus ponctuel marqué dans un cadre bayésien hiérarchique non supervisé où les sources sont détectées et leurs paramètres et les hyperparamètres du modèles sont estimés. Nous commencerons par introduire dans le paragraphe 2.1 la modélisation des galaxies et des données ainsi que les hypothèses qui doivent être faîtes afin de simplifier le problème d'estimation. Les paragraphes 2.2, 2.3 et 2.4 introduisent le modèle bayésien et le problème d'optimisation qu'il faut résoudre afin d'obtenir une estimation de la configuration de galaxies et des paramètres du modèle. Le processus d'échantillonnage et la structure de l'algorithme sont détaillés dans les paragraphes 2.5 et 2.6. Nous discuterons enfin des choix et hypothèses ainsi que des avantages et limitations de la méthode.

2.1 Modélisation du problème

Ce paragraphe a pour objectif d'introduire les paramètres du processus ponctuel marqué utilisé pour représenter la configuration de galaxies dans les données hyperspectrales, la modélisation de la PSF de l'instrument MUSE et le modèle d'observation adopté.

2.1.1 Modéliser la configuration de galaxies par un processus ponctuel marqué

La configuration de galaxies observée dans un cube de données MUSE est modélisée par un processus ponctuel marqué dont elle serait l'une des réalisations. Chaque point du processus ponctuel représente le centre d'une galaxie, des marques géométriques et d'intensité sont ensuite ajoutées pour transformer le point en objet.

Nous avons vu dans le paragraphe 1.5.3 que la projection d'une galaxie u_i dans le plan de l'observation MUSE pouvait être modélisée par une forme elliptique. Bien sûr, quelques galaxies spatialement étendues (et donc relativement proches) peuvent exhiber certaines caractéristiques spatiales qui ne sont pas en adéquation avec un modèle elliptique. Cependant dans le cadre de notre étude, la majorité des galaxies étant lointaines et faiblement résolues, cette approximation sera adoptée. Dans l'implémentation de l'algorithme nous avons représenté une ellipse par son centre (p_i, q_i), la taille des demi-axes a_i et b_i (sans distinction entre petit et grand axes) et son orientation α_i. La figure 2.1 représente les paramètres de l'ellipse. Comme il n'y a pas de distinction entre petit et grand axes dans notre représentation, l'orientation α_i définissant l'angle formé par l'horizontale et l'axe a_i (dans le sens trigonométrique) est comprise entre 0 et $\frac{\pi}{2}$. Les tailles minimales et maximales des demi-axes a_i et b_i sont largement dépendantes des données observées, pour des champs profonds les galaxies ont une extension faible spatiale, la taille maximale des demi-axes ne dépassera pas quelques pixels, tandis que pour des champs moins profond, on peut aisément imaginer que cette taille maximale peut excéder la dizaine de pixels. De plus la taille des demi-axes sera contrainte par le profil d'intensité choisi et les largeurs à mi-hauteur données à ce profil dans la direction des deux axes de l'ellipse.

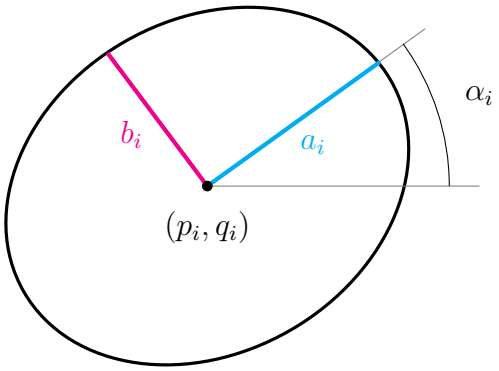


Figure 2.1 – Description d'un objet elliptique par sa position (p_i, q_i) et ses marques a_i, b_i, α_i.

En accord avec la littérature (voir paragraphe 1.5.3), nous avons choisi de modéliser la décroissance d'intensité centre-bord par un profil Sersic que l'on peut exprimer simplement par :

$$I(r) = I_0 \exp\left(-r^{\frac{1}{n}}\right) \tag{2.1}$$

où r est la distance au centre $O' = (p_i, q_i)$ de l'ellipse dans le repère elliptique $\tilde{\mathcal{R}} = (O', \vec{u}, \vec{v})$ tel qu'il est défini dans l'annexe C. Pour résumer, cette variable r peut s'écrire en fonction des

positions $(p, q) \in \mathcal{R} = (O, \overrightarrow{u}, \overrightarrow{v})$ de la façon suivante :

$$\begin{aligned} r^2 = &\frac{\cos^2(\alpha_i)}{\beta_1^2}(p-p_i)^2 + \frac{\sin^2(\alpha_i)}{\beta_1^2}(q-q_i)^2 + \frac{2\cos(\alpha_i)\sin(\alpha_i)}{\beta_1^2}(p-p_i)(q-q_i) \\ &+ \frac{\sin^2(\alpha_i)}{\beta_2^2}(p-p_i)^2 + \frac{\cos^2(\alpha_i)}{\beta_2^2}(q-q_i)^2 - \frac{2\cos(\alpha_i)\sin(\alpha_i)}{\beta_2^2}(p-p_i)(q-q_i) \end{aligned} \tag{2.2}$$

où $\mathcal{R}$ est le repère orthonormé correspondant à la grille de pixel des données, $\overrightarrow{u}$ et $\overrightarrow{v}$ sont les vecteurs directeurs des deux dimensions spatiales du cube de données. L'angle α_i correspond à la rotation de l'ellipse et les constantes β_1 et β_2 sont des paramètres d'échelle liés au repère elliptique $\tilde{\mathcal{R}}$ qui permettent d'ajuster la largeur à mi-hauteur du profil Sersic défini dans $\tilde{\mathcal{R}}$ le long des deux axes du support elliptique de l'objet u_i considéré. Le calcul de ces paramètres β_1 et β_2 en fonction des marques de l'objet u_i est détaillé dans l'annexe C. Le profil Sersic en deux dimensions associé à l'objet u_i présente une symétrie elliptique comme l'illustre la figure 2.2.

Finalement, concernant le processus ponctuel marqué, les marques à estimer pour chaque objet u_i de la configuration $\boldsymbol{u} = \{u_1, \cdots, u_{n(\boldsymbol{u})}\}$, où $n(\boldsymbol{u})$ désigne le nombre d'objets de la configuration $\boldsymbol{u}$, sont :

- les marques d'intensité : l'indice du profil Sersic n_i et les largeurs à mi-hauteur $FWHM_{a_i}$ et $FWHM_{b_i}$ au niveau des axes a_i et b_i du support elliptique.
- les marques géométriques de l'objet : α_i et la taille des demi-axes du support elliptique a_i et b_i qui est déduite à partir des deux largeurs à mi-hauteur et du profil Sersic choisi.

Il faut également ajouter la position (p_i, q_i) de l'objet. Finalemnt, il y a donc $6 \times n(\boldsymbol{u})$ paramètres à estimer pour définir la configuration d'objets détectés, où $n(\boldsymbol{u})$ est le nombre d'objets. Notons que ce décompte n'englobe pas la taille des demi-axes du support elliptique des objets puisqu'elle sont directement déduite à partir de la largeur à mi-hauteur du profil Sersic. Comme nous pouvons le voir dans l'annexe B, il est possible d'intégrer des informations concernant la localisation des objets directement dans la mesure du processus ponctuel de Poisson non homogène à l'aide d'une fonction d'intensité $\lambda(\cdot)$, voir sa définition (B.4). La façon dont est calculée cette fonction d'intensité est détaillée dans le chapitre 3, elle dépend fortement des données et du type d'objets que nous recherchons, dans un soucis de généralité nous considérerons une fonction d'intensité quelconque dans la suite de ce chapitre.

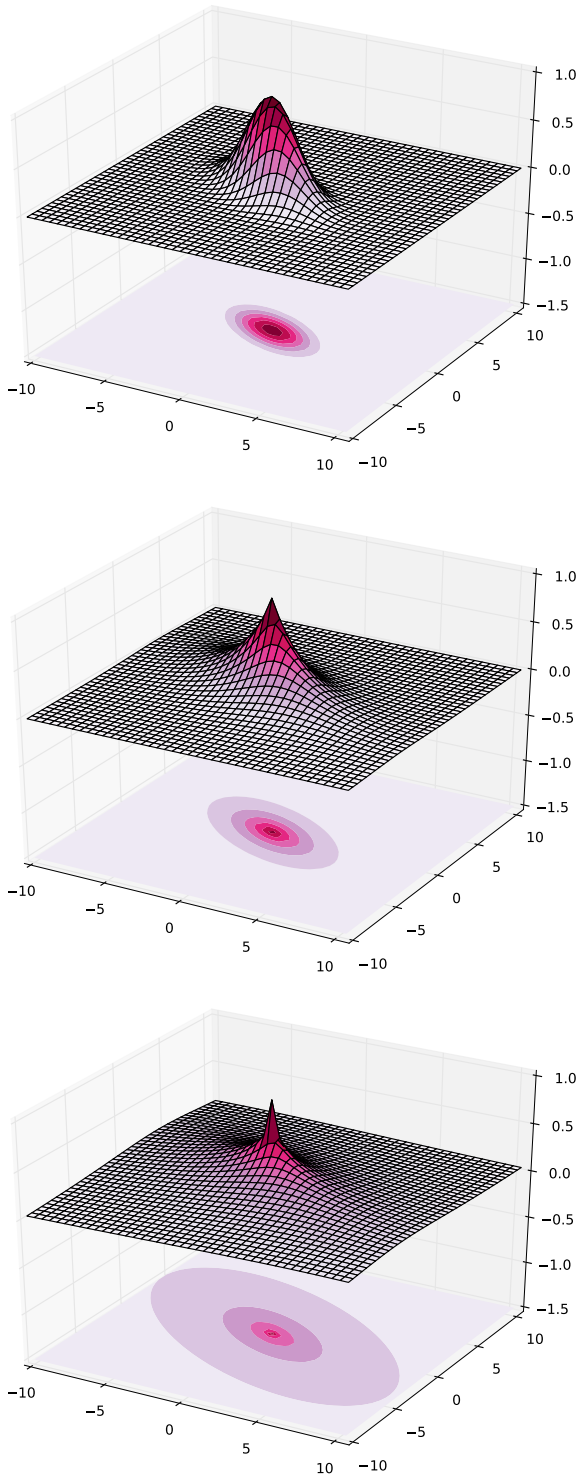


Figure 2.2 – Profils Sersic (non normalisés) et supports elliptiques associés définis selon les mêmes contraintes de centre, de largeurs à mi-hauteur et d'orientation pour différents indices n. Haut : $n = 0.5$, centre : $n = 1$ et bas = $n = 2$.

2.1.2 Observation d'une source en 3D

L'observation S_{obs} d'une source S avec l'instrument MUSE s'exprime donc de la façon suivante :

$$\mathrm{S}_{obs}(p,q,\lambda) = \sum_{z_p}\sum_{z_q}\sum_{\mu} L_\mu(\lambda) F_\mu(p-z_p, q-z_q) S(z_p, z_q, \mu) \tag{2.3}$$

Dans les données MUSE, nous modélisons une source S comme une fonction des trois dimensions du cube, cependant nous ajoutons un critère de séparabilité entre les dimensions spatiales et la dimension spectrale :

$$S(p,q,\lambda) = s(p,q)a(\lambda)$$

Cette hypothèse de séparabilité traduit le fait que nous considèrons qu'une galaxie garde la même forme spatiale et le même profil de décroissance d'intensité centre-bord, modélisés par s quelle que soit la longueur d'onde λ considérée. Le profil s est normalisé selon la norme ℓ_1 pour que le coefficient $a(\lambda)$ corresponde au flux lumineux émis par la source à la longueur d'onde λ. Les caractéristiques spectrales de la galaxie sont, quand à elles, décrites par le spectre $a = [a(1), \cdots, a(\lambda), \cdots, a(\Lambda)]$. Ainsi dans le cas d'une galaxie à spectre continu, toutes les composantes $a(\lambda)$ seront non nulles, tandis que dans le cas d'une galaxie avec une seule raie d'émission, seules quelques composantes adjacentes seront non nulles. Nous pouvons maintenant réécrire l'équation (2.3) de l'observation d'une source par l'instrument :

$$\mathrm{S}_{obs}(p,q,\lambda) = \sum_{\mu} L_\mu(\lambda)a(\mu) \sum_{z_p}\sum_{z_q} F_\mu(p-z_p, q-z_q) s(z_p, z_q) \tag{2.4}$$

Dans la suite de ce chapitre nous aurons besoin d'introduire une nouvelle hypothèse simplificatrice sur la FSF. En effet tenir compte de la variabilité spectrale de la FSF dans le processus de détection et de modélisation entraîne une complexité de calcul non négligeable étant données les dimensions du problème et le nombre d'objets observés dans un cube de données. Pour modéliser le profil s des objets, nous allons utiliser une FSF moyenne à la place des Λ expressions F_λ :

$$F(p,q) = \frac{1}{\Lambda}\sum_{\lambda=1}^{\Lambda} F_\lambda(p,q) \tag{2.5}$$

Nous obtenons alors une convolution spatiale indépendante de la longueur d'onde :

$$x(p,q) = \frac{1}{\|(F*s)\|_2}\sum_{z_p}\sum_{z_q} F(p-z_p, q-z_q) s(z_p, z_q) \tag{2.6}$$

Finalement x est la réponse spatiale moyenne normalisée (norme ℓ_2) de l'instrument à une source S. La normalisation ℓ_2 sera nécessaire dans la construction du modèle d'observation décrit dans le paragraphe 2.1.3. Il faut noter que si nous perdons l'information portée par la variation spectrale de la FSF dans ce modèle en utilisant une FSF moyenne F, nous exploiterons cette richesse dans une autre étape de la détection.

L'équation (2.4) peut donc se réécrire :

$$\begin{aligned}\mathrm{S}_{obs}(p,q,\lambda) &= \sum_{\mu} L_\mu(\lambda)a(\mu)\big\|F*s\big\|_2 x(p,q) \\ &= x(p,q)\big\|F*s\big\|_2 \sum_{\mu} L_\mu(\lambda)a(\mu) \\ &= x(p,q)\ w_\lambda\end{aligned} \tag{2.7}$$

où $w_\lambda = \|F * s\|_2 \sum_\mu L_\mu(\lambda) a(\mu)$ est l'intensité de la réponse de la source S à la longueur d'onde λ. Le vecteur $[w_1, \cdots, w_\Lambda]$ forme le spectre de la source S. Il faut noter que ce que nous appelons *spectre* dans ce manuscrit est en fait la convolution du spectre de la source observée avec la LSF de l'instrument, nous n'appliquerons pas de déconvolution dans ces travaux.

2.1.3 Le modèle d'observation

La configuration de galaxies à détecter est modélisée comme une réalisation d'un processus ponctuel marqué. Nous avons introduit dans le paragraphe 1.5.1 les hypothèses concernant la PSF de l'instrument et, dans le paragraphe précédent, le modèle adopté pour décrire une galaxie dans le cube de données hyperspectrales de MUSE. Nous allons maintenant expliquer ces données par un modèle qui décompose les observations en une contribution de la configuration de galaxie et la contribution du bruit décrit dans le paragraphe 1.5.2.

Avec l'hypothèse d'indépendance du bruit longueur d'onde par longueur d'onde formulée dans le paragraphe 1.5.2 et l'hypothèse que le profil spatial d'intensité des galaxies ne dépend par de la longueur d'onde considérée, nous allons pouvoir définir un modèle d'observation image par image (de taille $P \times Q$), *i.e.* pour une valeur fixée $\lambda \in [1, \ldots, \Lambda]$. Ce modèle peut être ensuite facilement étendu à un cube de donnée de taille $P \times Q \times \Lambda$.

Soit $\boldsymbol{y}_\lambda$ l'image vectorisée, $\boldsymbol{y}_\lambda$ est un vecteur de taille $M \times 1$ où $M = P \times Q$ est le nombre de pixels de l'image. Les sources présentes dans l'image sont modélisées par une configuration $\boldsymbol{u}$ de points auxquels sont ajoutées des marques (géométrie de la galaxie, profil d'intensité). En supposant que la configuration d'objets est connue dans ce paragraphe, les données $\boldsymbol{y}_\lambda$ sont modélisées de la façon suivante :

$$\boldsymbol{y}_\lambda = \mathbf{X}\boldsymbol{w}_\lambda + \mathbf{1}m_\lambda + \boldsymbol{\epsilon}_{Bg,\lambda}, \tag{2.8}$$

où m_λ est l'intensité moyenne du fond astrophysique, $\mathbf{1}$ est le vecteur unité de taille $M \times 1$, $\boldsymbol{\epsilon}_{Bg,\lambda}$ est le vecteur de taille $M \times 1$ représentant un bruit blanc gaussien centré tel que :

$$\boldsymbol{\epsilon}_{Bg,\lambda} \sim \mathcal{N}(0, \sigma_\lambda^2 \mathbf{I}_M) \tag{2.9}$$

où $\mathbf{I}_M$ est la matrice identité de taille $M \times M$. La configuration d'objets $\boldsymbol{u}$ est représentée par la matrice $\mathbf{X} = \left[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{n(\boldsymbol{u})}\right]$ qui est une matrice de taille $M \times n(\boldsymbol{u})$ dans laquelle chaque colonne $\boldsymbol{x}_j$ est l'image vectorisée de la composition d'une source u_j avec la FSF de l'instrument. Le nombre d'objets est noté $n(\boldsymbol{u})$ et $\boldsymbol{w}_\lambda = \left[w_{1,\lambda}, \ldots, w_{n(\boldsymbol{u}),\lambda}\right]^T \in \mathbb{R}^{n(\boldsymbol{u})}$ est le vecteur contenant les intensités de chaque source de $\mathbf{X}$ à la longueur d'onde λ donnée.

Il faut noter que même si le modèle monochromatique est défini pour une longueur d'onde λ donnée, l'observation à cette longueur d'onde λ est obtenue à l'aide de MUSE dont la PSF est en trois dimensions. L'impact de la FSF sur la réponse des sources est directement prise en compte dans le modèle des galaxies stockées dans la matrice $\mathbf{X}$ et elle ne dépend plus de la longueur d'onde (voir l'expression des $\boldsymbol{x}_i$, équation (2.6) dans le paragraphe 2.1.2). En revanche la corrélation induite par la LSF n'apparait pas explicitement dans dans le modèle d'observation. Elle est inclue directement dans l'estimation des intensités w_i de chaque objet (voir équation (2.7)) qui sera réalisée à partir du cube dans sa globalité.

Nous pouvons également noter que si nous souhaitons garder l'information de variabilité spectrale de la FSF dans le modèle d'observation, alors il faudra travailler avec une matrice $\mathbf{X}_\lambda = \left[\boldsymbol{x}_{1,\lambda} \ldots \boldsymbol{x}_{n(\boldsymbol{u}),\lambda}\right]$ par longueur d'onde λ où chaque réponse $\boldsymbol{x}_{i,\lambda}$ sera la convolution de la réponse s_i de la source avec la FSF définie à la longueur d'onde λ considéré F_λ. Il ne sera alors pas possible d'utiliser les formules de mise à jour récursive utilisées dans la procédure d'échantillonnage décrite dans le paragraphe 2.6.

2.2 Formulation bayésienne

Le modèle d'observation décrit par (2.8) dépend de nombreux paramètres inconnus (le nombre d'objets, leur position, leurs paramètres de forme, leur intensité, la moyenne et la variance du bruit) qu'il nous faut estimer. Il est nécessaire d'estimer la valeur de ces paramètres à l'aide de l'information portée par les observations.

2.2.1 Principe

Supposons que le modèle des observations $\boldsymbol{y}$ dépende d'un ensemble de paramètres $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)$:

$$\boldsymbol{y} \sim f(\boldsymbol{y}|\boldsymbol{\theta})$$

L'information portée par l'observation $\boldsymbol{y}$ est représentée par la densité $f(\boldsymbol{y}|\boldsymbol{\theta})$ que l'on appelle aussi vraisemblance, et qui est notée $\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{\theta})$. Dans la formulation bayésienne, les paramètres $\boldsymbol{\theta}$ ne sont pas considérés comme des quantités déterministes inconnues, mais comme des variables aléatoires que nous supposons distribuées suivant une distribution *a priori* notée $\pi(\boldsymbol{\theta})$, en l'absence d'observation .

Cette densité *a priori* résume la connaissance (ou l'absence de connaissance) que nous avons sur les paramètres $\boldsymbol{\theta}$ avant de faire l'observation. Les lois *a priori* peuvent être classées en deux grandes catégories :

- les lois dites *informatives* définies par l'utilisateur qui a une connaissance assez précise des valeurs prises par les paramètres $\boldsymbol{\theta}$,
- les lois dites *non informatives* qui ne favorisent pas de valeurs particulières pour les paramètres $\boldsymbol{\theta}$.

Le choix de l'*a priori* est important puisqu'il peut influer sur l'estimation finale des paramètres $\boldsymbol{\theta}$ selon le problème.

2.2.1.1 Densité a posteriori

La distribution *a posteriori* $p(\boldsymbol{\theta}|\boldsymbol{y})$ des paramètres $\boldsymbol{\theta}$ est obtenue par la formule de Bayes qui fait intervenir la distribution *a priori* $\pi(\boldsymbol{\theta})$ et la fonction de vraisemblance $f(\boldsymbol{y}|\boldsymbol{\theta})$:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

où $\int f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} = p(\boldsymbol{y})$ est l'évidence du modèle. Le calcul de cette densité *a posteriori* n'est pas toujours évident, notamment lorsqu'il est impossible d'obtenir une expression explicite de l'intégrale présente au dénominateur. Nous définirons dans la suite de ce manuscrit la densité *a posteriori* à une constante de normalisation près :

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$$

Nous verrons d'ailleurs par la suite qu'il est possible de mener l'estimation des paramètres $\boldsymbol{\theta}$ sans connaître la valeur de cette constante de normalisation. L'estimation des paramètres $\boldsymbol{\theta}$ se fera par maximisation de cette densité *a posteriori* afin d'obtenir les estimations au sens du maximum *a posteriori*.

2.2.1.2 *A priori* conjugué

Dans notre cas, la plupart du temps, l'utilisateur ne dispose d'aucune information *a priori* sur la valeur des paramètres, il est donc impossible de construire directement une densité *a priori*. Le calcul de la densité *a posteriori* des paramètres n'est pas toujours aisé. Cependant le

choix d'*a priori* conjugués permet de faciliter ce passage de la densité *a priori* $\pi(\boldsymbol{\theta})$ à la densité *a posteriori* $p(\boldsymbol{\theta}|\boldsymbol{y})$. On parle de loi *a priori* conjuguée lorsque $\pi(\boldsymbol{\theta})$ et $p(\boldsymbol{\theta}|\boldsymbol{y})$ font partie de la même famille de fonction. Une façon naturelle de choisir la famille de lois conjuguées est d'utiliser la forme de la fonction de vraisemblance. Lorsque cette distribution peut s'écrire sous la forme très générale $f(\boldsymbol{y}|\boldsymbol{\theta}) = a(\boldsymbol{y})b(\boldsymbol{\theta})\exp\left(\eta(\boldsymbol{\theta})T(\boldsymbol{y})\right)$, on dit que la distribution appartient à une famille exponentielle. Une famille exponentielle est une classe de distributions pour lesquelles il existe toujours une loi *a priori* conjuguée, qui, de plus, appartient à la même famille exponentielle. Nous renvoyons le lecteur à l'ouvrage de [Robert, 2006, p131] pour une description plus complète des lois *a priori* conjuguées pour les familles exponentielles usuelles dont nous donnons un extrait dans le tableau 2.1.

$f(\boldsymbol{y}\|\theta)$	$\pi(\theta)$	$p(\theta\|\boldsymbol{y})$
Normale $\mathcal{N}(\theta, \sigma^2)$	Normale $\mathcal{N}(\mu, \tau^2)$	$\mathcal{N}(\frac{\sigma^2\mu+\tau^2\boldsymbol{y}}{\sigma^2+\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2})$
Gamma $\mathcal{G}(\nu, \theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	$\mathcal{G}(\alpha+\nu, \beta+\boldsymbol{y})$
Normale $\mathcal{N}(\mu, 1/\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	$\mathcal{G}(\alpha+0.5, \beta+(\mu-\boldsymbol{y})^2/2)$

TABLEAU 2.1 – Quelques lois *a priori* conjuguées pour les familles exponentielles usuelles. Extrait du tableau [Robert, 2006, p131].

2.2.1.3 *A priori* non informatifs

Une densité *a priori* non informative est une loi qui n'apporte pas d'information dans l'inférence bayésienne. Le poids donné à la connaissance des paramètres $\boldsymbol{\theta}$ à estimer est réduit par rapport au poids donné aux observations. Ces loi *a priori* non informatives rendent la méthode d'estimation robuste puisqu'elle n'est pas biaisé par la subjectivité de l'*a priori*, elle est entièrement guidée par les données. La loi non informative la plus simple pour un paramètre θ_i défini sur un intervalle fermé Θ de taille t est l'*a priori* de Laplace, c'est une loi uniforme sur cet intervalle :

$$\theta_i \sim \frac{1}{t}$$

Jeffreys [1961] proposa une approche permettant de construire une loi *a priori* non informative en se basant sur l'information de Fisher $\mathbf{I}(\boldsymbol{\theta})$:

$$\mathbf{I}(\boldsymbol{\theta})_{i,j} = E\Big[-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(\boldsymbol{y}|\boldsymbol{\theta})\Big]$$

La matrice de Fisher mesure la quantité d'information disponible sur les paramètres $\boldsymbol{\theta}$ dans la distribution des données, plus $\mathbf{I}(\boldsymbol{\theta})$ sera importante plus l'observation apportera d'information sur $\boldsymbol{\theta}$. La stratégie de Jeffreys consiste à construire une loi *a priori* :

$$\pi(\boldsymbol{\theta}) \propto \left[\det\mathbf{I}(\boldsymbol{\theta})\right]^{1/2}$$

qui favorise les valeurs de $\boldsymbol{\theta}$ qui maximisent $\mathbf{I}(\boldsymbol{\theta})$. Cela traduit l'idée que les valeurs de $\boldsymbol{\theta}$ pour lesquelles $\mathbf{I}(\boldsymbol{\theta})$ est maximale doivent être *a priori* plus probables que les autres valeurs. Ainsi l'influence de la loi *a priori* sera moindre devant l'information délivrée par les données sur le paramètre $\boldsymbol{\theta}$. Les *a priori* de Jeffreys sont d'un grand intérêt puisqu'ils respectent la propriété

d'invariance par reparamétrisation :

$$\pi(\boldsymbol{\theta}) = \pi(f(\boldsymbol{\theta})) \left| \left(\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right) \right|$$

où $\left| \left(\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right) \right|$ est le déterminant de la matrice jacobienne de la transformation bijective f dont les éléments $\left(\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right)_{i,j}$ s'écrivent :

$$\left(\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right)_{i,j} = \frac{\partial f(\boldsymbol{\theta})_i}{\partial \boldsymbol{\theta}_j},$$

puisque l'on a :

$$\mathbf{I}(\boldsymbol{\theta}) = \mathbf{I}(f(\boldsymbol{\theta})) \left(\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} \right)^2$$

Si l'utilisation de loi *a priori* non informatives permet d'obtenir une estimation entièrement guidée par les données, il arrive fréquemment que la loi soit dite *impropre*, *i.e* $\int_\Theta \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} = +\infty$. Ces lois impropres peuvent faciliter le calcul de la loi *a posteriori*, il faut cependant faire attention à ce que la densité *a posteriori* obtenue soit bien définie, *i.e.* $\int_\Theta p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta} = 1$.

2.2.1.4 Un exemple de loi *a priori* : le *g-prior* (Zellner [1986])

Différents choix de paramétrisation des *a priori* existent dans la littérature, nous nous sommes intéressés à l'un d'entre eux : la distribution *g-prior* introduite par Zellner [1986] pour un modèle de régression linéaire gaussien :

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon \tag{2.10}$$

où $\boldsymbol{y} \in \mathbb{R}^n$ est le vecteur d'observation, $\mathbf{X} \in \mathbb{R}^{n \times d}$ est une matrice de design, $\boldsymbol{\beta} \in \mathbb{R}^d$ est le vecteur de coefficients des composantes et ϵ est un bruit gaussien tel que $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$. Zellner introduit un *a priori* gaussien pour le vecteur de coefficients, le *g-prior*, défini par :

$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, g^2\sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

Cet *a priori* est particulièrement simple, il ne dépend que d'un seul hyperparamètre g et sa structure de covariance, fixée par $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$, est ajustée directement à partir de la structure des données. Dans les travaux initiaux de Zellner [1986], l'auteur ne spécifie pas de méthodes particulières pour fixer g. Par la suite, d'autres auteurs ont utilisé et étudié ce *g-prior* et ont donné des pistes concernant le choix de g, voir par exemple les travaux de Kass and Wasserman [1996], Celeux et al. [2006], Celeux et al. [2012]. Le choix de l'hyperparamètre g peut se faire, de manière arbitraire en fixant g à une valeur donnée comme dans les travaux de Smith and Kohn [1996]. Il est aussi possible de fixer un *a priori* sur cet hyperparamètre afin de mener l'estimation dans un cadre entièrement bayésien. Nous obtenons alors un modèle hiérarchique où l'estimation du paramètre $\boldsymbol{\beta}$ dépend de l'estimation de l'hyperparamètre g. Un tel *a priori* est alors robuste puisqu'il est non informatif et l'estimation de $\boldsymbol{\beta}$ est entièrement guidée par les données.

Le *g-prior* est particulièrement intéressant puisque le paramètre g peut être interprété comme un rapport signal à bruit *a priori* sur le signal à détecter. En effet, l'expression du rapport signal

à bruit dans le cas du modèle de régression linéaire gaussien (2.10) s'écrit :

$$\begin{aligned}\frac{\|\mathbf{X}\boldsymbol{\beta}\|^2}{\|\epsilon\|^2} &= \frac{E\left[(\mathbf{X}\boldsymbol{\beta})^T(\mathbf{X}\boldsymbol{\beta})\right]}{E\left[\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}\right]} \\ &= \frac{\mathrm{Tr}\left\{E\left[\boldsymbol{\beta}\boldsymbol{\beta}^T\right]\mathbf{X}^T\mathbf{X}\right\}}{\sigma^2\,\mathrm{Tr}\left\{I_n\right\}} \\ &= \frac{\mathrm{Tr}\left\{g^2\sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\right\}}{\sigma^2\,\mathrm{Tr}\left\{I_n\right\}} \\ &= g^2\end{aligned}$$

où l'on retrouve l'hyperparamètre g^2 de la loi *a priori* du vecteur de coefficient $\boldsymbol{\beta}$.

2.2.2 Vraisemblance

D'après le modèle d'observation décrit par (2.8) et (2.9) la vraisemblance monochromatique s'écrit :

$$f(\boldsymbol{y}_\lambda|\boldsymbol{u},\boldsymbol{w}_\lambda,m_\lambda,\sigma_\lambda^2) = \left(\frac{1}{2\pi\sigma_\lambda^2}\right)^{\frac{M}{2}}\exp\left(-\frac{(\boldsymbol{y}_\lambda-\mathbf{X}\boldsymbol{w}_\lambda-\mathbf{1}m_\lambda)^T(\boldsymbol{y}_\lambda-\mathbf{X}\boldsymbol{w}_\lambda-\mathbf{1}m_\lambda)}{2\sigma_\lambda^2}\right) \tag{2.11}$$

En considérant que le bruit $\boldsymbol{\epsilon}_{Bg,\lambda}$ est indépendant selon les longueurs d'onde, il possible d'écrire la vraisemblance globale des données sur les Λ longueurs d'onde comme le produit des vraisemblances monochromatiques :

$$f(\mathbf{Y}|\boldsymbol{u},\mathbf{W},\boldsymbol{m},\boldsymbol{\sigma^2}) = \prod_{\lambda=1}^{\Lambda} f(\boldsymbol{y}_\lambda|\boldsymbol{u},\boldsymbol{w}_\lambda,m_\lambda,\sigma_\lambda^2)$$

où $\mathbf{Y} = [\boldsymbol{y}_1,...,\boldsymbol{y}_\Lambda]$ est une matrice de taille $M\times\Lambda$ qui contient les Λ images vectorisées et $\mathbf{W} = [\boldsymbol{w}_1,...,\boldsymbol{w}_\Lambda]$ est une matrice de taille $n(\boldsymbol{u})\times\Lambda$. Cette matrice $\mathbf{W}$ possède une double interprétation :

$$\mathbf{W} = \begin{pmatrix} w_{1,1} & \cdots & w_{1,\lambda} & \cdots & w_{1,\Lambda} \\ \vdots & & \vdots & & \vdots \\ w_{i,1} & \cdots & w_{i,\lambda} & \cdots & w_{i,\Lambda} \\ \vdots & & \vdots & & \vdots \\ w_{n(\boldsymbol{u}),1} & \cdots & w_{n(\boldsymbol{u}),\lambda} & \cdots & w_{n(\boldsymbol{u}),\Lambda} \end{pmatrix} \tag{2.12}$$

où :

- chaque colonne $\boldsymbol{w}_\lambda = \left[w_{1,\lambda},\ldots,w_{n(\boldsymbol{u}),\lambda}\right]^T$ est un vecteur contenant l'intensité des $n(\boldsymbol{u})$ objets à la longueur d'onde $\lambda\in[\![1,\Lambda]\!]$,
- chaque ligne $\boldsymbol{w}_i = [w_{i,1},\ldots,w_{i,\lambda},\ldots w_{i,\Lambda}]$ est le spectre du $i^{\text{ème}}$ objet, $i\in[\![1,n(\boldsymbol{u})]\!]$.

Nous définissons également les vecteurs de taille $1\times\Lambda$, $\boldsymbol{m} = [m_1,...,m_\Lambda]$ et $\boldsymbol{\sigma^2} = \left[\sigma_1^2,...,\sigma_\Lambda^2\right]$, qui correspondent à la moyenne et à la variance du bruit définies pour chaque longueur d'onde.

2.3 Les *a priori* sur les paramètres du modèle

Le modèle d'observation décrit par (2.8) dépend de nombreux paramètres inconnus dont il va falloir définir la densité *a priori*. Les paramètres concernés sont :

- le nombre d'objets $n(\boldsymbol{u})$ ainsi que leur positions (p_i,q_i) et leurs paramètres de forme (a_i,b_i,α_i,n_i) ; tous ces paramètres représentent le processus ponctuel marqué et permettent de construire la matrice de configuration $\mathbf{X}$,
- la moyenne m_λ et la variance du bruit de fond σ_λ^2 à chaque longueur d'onde λ,
- l'intensité des objets $\boldsymbol{w}_\lambda$ à chaque longueur d'onde λ.

2.3.1 Paramètres du bruit de fond

Puisque nous ne disposons d'aucune information *a priori* sur l'intensité du bruit de fond ou sur sa variance, nous choisissons un *a priori* non informatif de Jeffreys pour le couple de paramètre $(m_\lambda, \sigma_\lambda)$. Cette densité aura la même forme pour tous les couples $(m_\lambda, \sigma_\lambda)$, pour tout $1 \leqslant \lambda \leqslant \Lambda$. La règle de calcul de la loi *a priori* de Jeffreys [1961] appliquée au couple de paramètres $(m_\lambda, \sigma_\lambda^2)$ est donc :

$$\pi(m_\lambda, \sigma_\lambda^2) \propto \left[\det(\mathbf{I}(m_\lambda, \sigma_\lambda^2)) \right]^{\frac{1}{2}}$$

où $\mathbf{I}(m_\lambda, \sigma_\lambda^2)$ est la matrice de Fisher. Posons $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \mathbf{X}\boldsymbol{w}$ pour simplifier les notations, la matrice de Fisher s'écrit alors :

$$\mathbf{I}(m_\lambda, \sigma_\lambda^2) = -E\left[\begin{pmatrix} \frac{M}{\sigma_\lambda^2} & \frac{2(1^T\tilde{\boldsymbol{y}} - Mm_\lambda)}{\sigma_\lambda^3} \\ \frac{2(1^T\tilde{\boldsymbol{y}} - Mm_\lambda)}{\sigma_\lambda^3} & \frac{3(\tilde{\boldsymbol{y}} - 1m_\lambda)^T(\tilde{\boldsymbol{y}} - 1m_\lambda)}{\sigma_\lambda^4} - \frac{M}{\sigma_\lambda^2} \end{pmatrix} \right]$$

$$= \begin{pmatrix} \frac{M}{\sigma_\lambda^2} & 0 \\ 0 & \frac{2M}{\sigma_\lambda^2} \end{pmatrix}$$

et donc la densité *a priori* s'écrit :

$$\pi(m_\lambda, \sigma_\lambda^2) \propto \frac{1}{\sigma_\lambda^2} \mathbb{1}_{]0,+\infty[}(\sigma_\lambda^2) \mathbb{1}_{\mathbb{R}}(m_\lambda) \tag{2.13}$$

Il faut noter que cette loi *a priori* est impropre puisqu'il n'est pas possible de le normaliser : $\int \pi(m_\lambda, \sigma_\lambda^2) dm_\lambda d\sigma_\lambda = +\infty$, cependant, nous verrons dans le paragraphe 2.4 que la densité *a posteriori* résultante est bien définie. Notons aussi que le domaine d'existence du paramètre m_λ n'est pas restreint, $m_\lambda \in \mathbb{R}$, car du fait des opérations de soustraction de ciel, certains pixels présentent des valeurs négatives, nous ne pouvons donc garantir que la moyenne du bruit de fond soit positive. Ceci est vrai, en particulier aux longueurs d'onde correspondantes aux raies d'émission du ciel, car du fait du sous échantillonnage de la LSF de MUSE, son estimation n'est pas parfaite et entraine des erreurs de modélisation des raies du ciel qui se répercutent lors de la soustraction.

2.3.2 Intensité des objets

Concernant les intensités des objets, deux *a priori* différents ont été envisagés. A l'origine, dans le modèle présenté dans les travaux de Chatelain et al. [2011] l'*a priori* sur les intensités des objets était trop régularisant, nous verrons que ce choix pénalise la détection des objets les moins brillants. L'objectif de cette thèse étant de rendre la méthode capable de détecter non seulement les objets brillants mais aussi les objets de plus faible intensité, les plus lointains, nous avons opté pour un *a priori* non informatif qui ne favorise pas en amont les objets de forte intensité.

Dans les travaux précédents de Chatelain et al. [2011], un *g-prior* était utilisé pour le vecteur d'intensité $\boldsymbol{w}_\lambda \in \mathbb{R}^{n(\boldsymbol{u})}$:

$$p(\boldsymbol{w}_\lambda | \boldsymbol{u}, \sigma_\lambda^2, g_\lambda^2) \sim \mathcal{N}\left(0, g_\lambda^2 \sigma_\lambda^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right)$$

où l'hyperparamètre g_λ^2 représente un rapport signal à bruit *a priori* de la scène observée. L'utilisation de cet *a priori* introduit une régularisation sur les intensités, ce qui permet à l'algorithme d'accepter des objets uniquement aux endroits où les intensités sont suffisamment élevées. Cet

a priori est beaucoup trop contraignant pour les données MUSE qui contiennent des sources dont les intensités présentent des grandes dynamiques inter-objets, une valeur trop élevée de l'hyperparamètre g_λ^2 pénalise la détection des galaxies les moins brillantes. Cet *a priori* sera, en revanche, tout à fait adapté si tous les objets ont des intensités similaires.

En l'absence d'information sur les intensités des objets, nous utilisons un *a priori* non informatif de Jeffreys :

$$p(\boldsymbol{w}_\lambda|\boldsymbol{u}) \propto \mathbb{1}_{\mathbb{R}^{n(\boldsymbol{u})}}(\boldsymbol{w}_\lambda). \tag{2.14}$$

L'utilisation de cette densité entraîne une suppression complète de la régularisation des intensités dans le modèle et dans la distribution *a posteriori* définie par la suite (équation (2.23)). En l'absence de régularisation sur les intensités des objets dans la distribution *a posteriori*, les données ont une influence forte sur l'estimation au sens du maximum *a posteriori*. En l'absence de régularisation sur les intensités des objets, seules les données influent sur l'estimation des objets proposés. Ainsi, un pixel de bruit isolé de forte intensité risque d'être vu par l'algorithme d'optimisation comme une contribution à modéliser. Ajouter un petit objet à cet emplacement permet d'expliquer cette contribution dans les données, mais d'un point de vue de la détection d'objets, ce n'est évidemment pas satisfaisant. Nous verrons qu'avec cet *a priori* non informatif, il faudra ajouter un autre critère de régularisation sur la configuration d'objet.

Il faut noter que la positivité des intensités des objets n'est pas explicitement définie ici dans cet *a priori*. Si $\boldsymbol{w}_\lambda$ était défini sur $\mathbb{R}^{n(\boldsymbol{u})+}$, alors la distribution *a posteriori* du vecteur d'intensité $\boldsymbol{w}_\lambda$ serait une loi gaussienne multivariée tronquée. La constante de normalisation d'une telle distribution n'a pas de forme explicite, or nous en aurions besoin lors de la marginalisation du paramètre $\boldsymbol{w}$ dans la distribution *a posteriori*. D'un point de vue calculatoire, le choix de ne pas tronquer à $\mathbb{R}^{n(\boldsymbol{u})+}$ est donc avantageux. De plus, nous savons qu'avec les opérations de soustraction de ciel appliquées aux données, les intensités peuvent parfois être négatives, c'est pourquoi nous ne pouvons garantir une positivité du spectre d'un objet sur l'ensemble des longueurs d'onde. Nous utiliserons donc plutôt un critère global sur le spectre en s'intéressant notamment à la statistique de la valeur maximum du spectre des objets.

2.3.3 Configuration d'objets

La configuration d'objets, représentée par $\boldsymbol{u}$ dans le formalisme du processus ponctuel marqué ou par $\mathbf{X}$ dans le formalisme matriciel, est décrite par une multitude de paramètres à définir. Nous avons traité dans le paragraphe précédent des intensités $\boldsymbol{w}$ de ces objets (et donc implicitement de leur spectre), nous allons nous intéresser ici à la distribution des objets et aux contraintes appliquées sur les relations inter-objets.

2.3.3.1 *A priori* sur la mesure de référence du processus ponctuel

La mesure de référence du processus ponctuel donnée dans l'équation (B.8) dépend d'un paramètre d'intensité β. En effet le nombre d'objets $n(\boldsymbol{u})$ de la configuration $\boldsymbol{u}$ est distribué selon une loi de Poisson de moyenne β. La densité du processus de référence définie par rapport à la mesure normalisée $\pi_{\nu'}$ du processus ponctuel de Poisson s'écrit :

$$\begin{aligned} f(\boldsymbol{u}|\beta) &= \beta^{n(\boldsymbol{u})}\exp(-(\beta-1)), \\ &\propto \beta^{n(\boldsymbol{u})}e^{-\beta}. \end{aligned} \tag{2.15}$$

Dans le cadre Bayesien, le paramètre β est maintenant considéré comme une variable aléatoire. Dans le contexte de MUSE, il n'y a aucune information *a priori* concernant cette intensité β, nous choisissons un *a priori* conjugué mais en fixant les paramètres de manière à obtenir un *a*

priori vague, *i.e.* avec une très grande variance, et qui ne favorise donc pas particulièrement une valeur. Nous choisissons ici une loi conjuguée Gamma $\mathcal{G}(a, b)$ pour l'hyperparamètre β :

$$p(\beta) = \frac{\beta^{a-1} e^{-\beta/b}}{\Gamma(a) b^a}, \qquad \forall \beta \in \mathbb{R}^+, \tag{2.16}$$

où $\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt$ est la fonction Gamma. Afin de rendre la densité suffisamment vague, les hyperparamètres a et b sont fixés aux valeurs $a = 1$ et $b = 10^3$. En appliquant la formule de Bayes, nous obtenons alors :

$$f(\boldsymbol{u}, \beta) = f(\boldsymbol{u}|\beta) p(\beta) \propto \beta^{n(\boldsymbol{u})+a-1} \exp\left(-\frac{1+b}{b}\beta\right). \tag{2.17}$$

L'équation (2.17) montre que la densité *a posteriori* conditionnelle de β connaissant la configuration $\boldsymbol{u}$ est une loi Gamma : $\beta \sim \mathcal{G}\left(n(\boldsymbol{u}) + a, b/(1+b)\right)$. Cet hyperparamètre β constituant un paramètre de nuisance (nous ne cherchons pas à l'estimer), il est possible de marginaliser $f(\boldsymbol{u}, \beta)$, ce qui induit la densité *a priori* suivante pour la configuration de points :

$$f(\boldsymbol{u}) \propto \Gamma\left(n(\boldsymbol{u}) + 1\right) q^{n(\boldsymbol{u})+1}, \tag{2.18}$$

où $q = b/(b+1)$. Notons que le processus défini par cette densité est un cas particulier d'un processus ponctuel binomial négatif décrit par Diggle and Milne [1983]. Le nombre de points dans chaque ensemble compact est distribué selon une loi binomiale négative. De plus le choix d'un *a priori* vague (2.16) pour l'hyperparamètre β assure d'obtenir un hyperparamètre q proche de 1.

2.3.3.2 *A priori* sur la configuration d'objets

Dans le paragraphe précédent nous avons introduit un *a priori* sur la mesure de référence du processus ponctuel, *i.e.* sur les positions et sur le nombre de galaxies dans l'image. Nous allons maintenant ajouter un *a priori* sur les marques du processus ponctuel marqué et sur les relations entre les objets.

Nous rappelons que les marques définissant un objet u_i sont : la taille des demi-axes (a_i, b_i), l'orientation de l'ellipse par rapport à l'horizontale α_i et l'indice du profil Sersic n_i. Puisque nous ne disposons d'aucune information extérieure sur les objets, nous utilisons des *a priori* uniformes sur le domaine d'existence de ces marques :

- $a_i \sim \mathcal{U}(r_{min}, r_{max})$
- $b_i \sim \mathcal{U}(r_{min}, r_{max})$
- $\alpha_i \sim \mathcal{U}(0, \frac{\pi}{2})$
- $n_i \sim \mathcal{U}\{0.5, 1, 2\}$

Nous définissons alors un *a priori* global pour toutes les marques résumées par :

$$\pi(a_i, b_i, \alpha_i, n_i) = \frac{1}{(r_{max} - r_{min})^2} \mathbb{1}_{[r_{min}, r_{max}]^2}(a_i, b_i) \frac{2}{\pi} \mathbb{1}_{[0, \frac{\pi}{2}]}(\alpha_i) \frac{1}{3} \mathbb{1}_{\{0.5,1,2\}}[n_i] \tag{2.19}$$

Afin d'alléger les notations, cet *a priori* sera modélisé comme un terme de pénalisation *hard-core* $h_1(\boldsymbol{u})$ interdisant toutes les configurations contenant au moins un objet ne respectant pas les critères de forme et d'intensité définis ci-dessus que l'on peut résumer par :

$$h_1(\boldsymbol{u}) = \begin{cases} 0 \text{ si } \exists u_i \in \boldsymbol{u} \text{ t.q.} & a_i \notin [r_{min}, r_{max}] \\ & \text{ou } b_i \notin [r_{min}, r_{max}] \\ & \text{ou } \alpha_i \notin [0, \frac{\pi}{2}] \\ & \text{ou } n_i \notin \{0.5, 1, 2\} \\ 1 \text{ sinon} \end{cases}, \tag{2.20}$$

si bien que le produit des *a priori* sur les objets soit défini comme :

$$\prod_{i=1}^{n(\boldsymbol{u})} \pi(a_i, b_i, \alpha_i, n_i) \propto h_1(\boldsymbol{u})$$

Afin d'éviter des détections multiples, c'est-à-dire la détection d'une galaxie par plusieurs objets, nous allons introduire un critère de pénalisation des recouvrement entre objets dans la densité *a priori*. Notons $r(u_i, u_j)$ le ratio de recouvrement entre les objets u_i et u_j :

$$r(u_i, u_j) = \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$$

où $\boldsymbol{x}_i$ est la réponse normalisée ℓ_2 de la source u_i et $\langle \cdot, \cdot \rangle$ désigne le produit scalaire euclidien. Le ratio $r(u_i, u_j) \in [0, 1]$ nous permet de construire une seconde pénalisation à travers la densité $h_2(\boldsymbol{u})$ définie par rapport au processus de référence :

$$h_2(\boldsymbol{u}) = \begin{cases} 0 \text{ s'il existe } i \neq j \text{ tel que } r(u_i, u_j) > t, \\ 1 \text{ sinon,} \end{cases} \tag{2.21}$$

où $t \in [0, 1]$ est le ratio maximal de recouvrement autorisé. On interdit donc à travers la pénalisation $h_2(\boldsymbol{u})$ toutes les configurations d'objets contenant au moins deux objets se recouvrant avec un ratio supérieur à t.

Le produit scalaire qui apparaît dans l'expression de $r(u_i, u_j)$ est la métrique naturelle pour mesurer la colinéarité des colonnes de la matrice $\mathbf{X}$, *i.e.* la ressemblance entre la réponse des objets de la configuration. Interdire un taux de recouvrement proche de 1 permet notamment de contrôler l'inversibilité et le conditionnement de la matrice de Gram $\mathbf{X}^T\mathbf{X}$ qui apparaît dans l'expression de la densité *a posteriori* (voir paragraphe 2.4).

Le paramètre t correspond au seuil au dessus duquel deux objets seront considérés comme non résolus ou trop similaires pour être considérés comme la détection de deux galaxies distinctes. Ce taux de recouvrement peut être vu comme un hyperparamètre du modèle bayésien hiérarchique et il est nécessaire ici de le fixer à partir des connaissances que nous avons de l'instrument MUSE. La résolution spatiale de MUSE est donnée par le critère de Rayleigh. Nous fixerons donc la valeur de t en se plaçant à la limite de résolution de la FSF de MUSE, définie par sa largeur à mi-hauteur, voir équation (1.14). Cette valeur de t devra être ajustée à chaque nouveau cube de données puisque la FSF de MUSE varie.

Il faut noter que si la forme de tous les objets est un disque de taille fixe, tester le produit scalaire des deux objets revient à tester la distance entre les centres des objets. Ici les objets ont des formes elliptiques dont la taille, la forme et l'orientation peuvent changer, on ne peut donc pas uniquement se baser sur la distance entre les centres, c'est pourquoi nous avons choisi le produit scalaire.

2.3.3.3 *A priori* global sur le processus ponctuel marqué

Finalement les deux contraintes $h_1(\boldsymbol{u})$ et $h_2(\boldsymbol{u})$ appliquées sur la configuration d'objets peut se résumer en un seul terme de pénalisation *hard-core* $h(\boldsymbol{u})$ se résumant par :

$$h(\boldsymbol{u}) = h_1(\boldsymbol{u}) h_2(\boldsymbol{u}) \tag{2.22}$$

La densité *a priori* globale du processus ponctuel marqué $p(\boldsymbol{u})$ définie par rapport à la mesure du processus de Poisson de référence $\pi_{\nu'}$ intègre la densité du processus ponctuel $f(\boldsymbol{u})$ et les contraintes définies par l'*a priori* $h(\boldsymbol{u})$ sur la configuration d'objet dans l'équation (2.22) :

$$p(\boldsymbol{u}) \propto f(\boldsymbol{u}) h(\boldsymbol{u})$$

2.4 Densité *a posteriori*

Maintenant que tous les *a priori* sur les paramètres inconnus sont définis ainsi que le modèle d'observation, nous allons pouvoir définir la densité conjointe *a posteriori* des paramètres connaissant les données. Cette densité est complexe, elle dépend d'un grand nombre de paramètres. Il est impossible d'avoir la forme explicite de l'estimateur au sens du maximum *a posteriori* (MAP) de ces paramètres directement à partir de l'expression de la densité *a posteriori*. Nous allons devoir échantillonner la densité *a posteriori* et extraire le meilleur jeu de paramètres qui maximise la densité *a posteriori* jointe.

2.4.1 Expression de la densité *a posteriori* jointe des paramètres $\boldsymbol{u}$, $\boldsymbol{w}$, $\boldsymbol{m}$ et $\boldsymbol{\sigma^2}$

D'après le modèle d'observation monochromatique (2.8) et les *a priori* choisis, on peut écrire la densité conjointe *a posteriori* des paramètres $\boldsymbol{w}_\lambda$, m_λ et σ^2_λ conditionnellement aux observations $\boldsymbol{y}_\lambda$ et à la configuration d'objets $\boldsymbol{u}$. Comme nous avons fixé la configuration d'objets $\boldsymbol{u}$ identique pour toutes les longueurs d'onde, sa loi *a priori* sera introduite dans la densité *a posteriori* globale. La densité jointe *a posteriori* des paramètres $\boldsymbol{w}_\lambda$, m_λ et σ^2_λ pour une longueur d'onde λ fixée conditionnellement à la configuration $\boldsymbol{u}$ s'écrit :

$$\begin{aligned} p(\boldsymbol{w}_\lambda, m_\lambda, \sigma^2_\lambda | \boldsymbol{y}_\lambda, \boldsymbol{u}) \propto\ & p(\boldsymbol{y}_\lambda | \boldsymbol{w}_\lambda, m_\lambda, \sigma^2_\lambda, \boldsymbol{u}) p(m_\lambda, \sigma^2_\lambda) p(\boldsymbol{w}_\lambda | \boldsymbol{u}) \\ & \propto \left(\frac{1}{2\pi\sigma^2_\lambda}\right)^{\frac{M}{2}} \times \exp\left(-\frac{(\boldsymbol{y}_\lambda - \mathbf{X}_\lambda \boldsymbol{w}_\lambda - \mathbf{1} m_\lambda)^T (\boldsymbol{y}_\lambda - \mathbf{X}_\lambda \boldsymbol{w}_\lambda - \mathbf{1} m_\lambda)}{2\sigma^2_\lambda}\right) \\ & \quad \times \frac{1}{\sigma^2_\lambda} \mathbb{1}_{]0,+\infty[}(\sigma^2_\lambda) \mathbb{1}_{\mathbb{R}^{n(\boldsymbol{u})}}(\boldsymbol{w}_\lambda). \end{aligned} \tag{2.23}$$

Et donc la densité *a posteriori* globale s'écrit :

$$\begin{aligned} p(\boldsymbol{u}, \mathbf{W}, \boldsymbol{m}, \boldsymbol{\sigma^2} | \mathbf{Y}) \propto & \prod_{\lambda=1}^{\Lambda} \Big\{ p(\boldsymbol{w}_\lambda, m_\lambda, \sigma^2_\lambda | \boldsymbol{u}, \boldsymbol{y}_\lambda) \Big\} p(\boldsymbol{u}) \\ & \propto \prod_{\lambda=1}^{\Lambda} \Big\{ \left(\frac{1}{2\pi\sigma^2_\lambda}\right)^{\frac{M}{2}} \exp\left(-\frac{(\boldsymbol{y}_\lambda - \mathbf{X}_\lambda \boldsymbol{w}_\lambda - \mathbf{1} m_\lambda)^T (\boldsymbol{y}_\lambda - \mathbf{X}_\lambda \boldsymbol{w}_\lambda - \mathbf{1} m_\lambda)}{2\sigma^2_\lambda}\right) \\ & \quad \times \frac{1}{\sigma^2_\lambda} \mathbb{1}_{]0,+\infty[}(\sigma^2_\lambda) \mathbb{1}_{\mathbb{R}^{n(\boldsymbol{u})}}(\boldsymbol{w}_\lambda) \Big\} \\ & \quad \times \Gamma\left(n(\boldsymbol{u}) + 1\right) \times q^{n(\boldsymbol{u})+1} h(\boldsymbol{u}). \end{aligned} \tag{2.24}$$

2.4.2 Marginalisation des paramètres de nuisance

A partir de cette densité *a posteriori* (2.24) nous allons échantillonner l'ensemble des paramètres inconnus et retenir le meilleur jeu de paramètre au sens du maximum *a posteriori*. Il faut donc échantillonner $6 \times n(\boldsymbol{u}) + n(\boldsymbol{u}) \times \Lambda + 2 \times \Lambda$ paramètres pour les marques du processus ponctuel marqué, les intensités des $n(\boldsymbol{u})$ objets sur les Λ longueurs d'onde et les valeurs de moyenne et variance du bruit de fond sur les Λ longueurs d'onde. Pour un cube de données MUSE de taille standard, $\Lambda \simeq 3600$ et $n(\boldsymbol{u})$ est de l'ordre de quelques centaines, la dimension du problème est donc considérable. De plus un trop grand nombre de paramètres est discriminant pour la convergence d'un échantillonneur de type Metropolis-Hastings(-Green), puisqu'il suffit que l'un de ces paramètres ait été échantillonné à une valeur extreme pour que l'étape d'acceptation-rejet du mouvement conduise à un rejet du mouvement complet, et donc à un rejet de tous

les paramètres. Plus le nombre de paramètres à échantillonner simultanément sera élevé, plus la proportion de mouvements rejetés sera importante. Il est possible de marginaliser la densité (2.24) selon certains paramètres (par exemple les intensités des objets) afin de réduire le nombre de paramètres à échantillonner. En développant le terme dans l'exponentielle, il est possible de faire apparaître la densité *a posteriori* du vecteur d'intensité $\boldsymbol{w}_\lambda$ pour chaque longueur d'onde λ fixée conditionnellement aux autres paramètres $m_\lambda, \sigma_\lambda^2, \boldsymbol{u}$ et $\boldsymbol{y}_\lambda$. Cette densité conditionnelle *a posteriori* est gaussienne de moyenne $\boldsymbol{\mu}_\lambda$ et de matrice de covariance $\boldsymbol{\Sigma}_{\boldsymbol{w}_\lambda}$ exprimée par :

$$\begin{aligned}\boldsymbol{\mu}_\lambda &= (\mathbf{X}_\lambda^T\mathbf{X}_\lambda)^{-1}\left[\mathbf{X}_\lambda^T(\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda)\right],\\ \boldsymbol{\Sigma}_{\boldsymbol{w}_\lambda} &= \sigma_\lambda^2\left(\mathbf{X}_\lambda^T\mathbf{X}_\lambda\right)^{-1}.\end{aligned}$$

En marginalisant selon ce paramètre la densité *a posteriori* définie dans (2.24) devient, après calcul (voir le calcul détaillé dans l'annexe I.1) :

$$\begin{aligned}p(\boldsymbol{u}, \boldsymbol{m}, \boldsymbol{\sigma^2}|\mathbf{Y}) &\propto \prod_{\lambda=1}^{\Lambda}\Big\{p(m_\lambda, \sigma_\lambda^2|\boldsymbol{u}, \boldsymbol{y}_\lambda)\Big\}p(\boldsymbol{u})\\ &\propto \prod_{\lambda=1}^{\Lambda}\Big\{\left(\frac{1}{2\pi\sigma_\lambda^2}\right)^{\frac{M-n(\boldsymbol{u})}{2}} e^{-\frac{(\boldsymbol{y}_\lambda-\mathbf{1}m_\lambda)^T\mathbf{V}_\lambda(\boldsymbol{y}_\lambda-\mathbf{1}m_\lambda)}{2\sigma_\lambda^2}}\left|\left(\mathbf{X}_\lambda^T\mathbf{X}_\lambda\right)^{-1}\right|^{\frac{1}{2}} \times \frac{1}{\sigma_\lambda^2}\mathbb{1}_{]0,+\infty[}(\sigma_\lambda^2)\Big\}\\ &\quad \times \Gamma\left(n(\boldsymbol{u})+1\right) \times q^{n(\boldsymbol{u})+1}h(\boldsymbol{u}).\end{aligned} \tag{2.25}$$

avec $\mathbf{V}_\lambda = \mathbf{I}_M - \mathbf{X}_\lambda(\mathbf{X}_\lambda^T\mathbf{X}_\lambda)^{-1}\mathbf{X}_\lambda^T$ qui définit la projection sur le sous espace du bruit. Notons que l'inversibilité de la matrice de Gram $\mathbf{X}_\lambda^T\mathbf{X}_\lambda$ est assurée par notre choix d'*a priori* sur la configuration d'objets (2.21). Notons également que même si nous avons choisi des *a priori* non informatifs impropres (2.13) et (2.14) la densité *a posteriori* des paramètres est bien définie.

2.5 Echantillonnage des paramètres inconnus

Bien que la densité *a posteriori* des paramètres soit bien définie (équation (2.25)), il est impossible d'extraire analytiquement l'estimateur au sens du MAP des paramètres inconnus. Nous allons donc utiliser une méthode RJMCMC pour générer des échantillons de ces paramètres dont la loi approche la densité *a posteriori*. Les différents algorithmes d'échantillonnage utilisés par la suite sont décrit dans l'annexe B. Le processus d'échantillonnage est décrit dans les paragraphes suivants.

2.5.1 Echantillonnage des paramètres du bruit

Etant donnée la distribution *a posteriori* marginalisée (2.25), les distributions conditionnelles des paramètres du bruit m_λ et σ_λ^2 peuvent être déduites directement. Ces distributions sont bien définies et sont connues, des échantillons de ces distributions peuvent être facilement générés grâce à différents langages ou logiciels de calcul pour construire les chaînes de Markov $\{m_\lambda^{(k)}\}_k$ et $\{\sigma_\lambda^{2(k)}\}_k$, pour tout $1 \leqslant \lambda \leqslant \Lambda$. Posons :

$$\begin{aligned}\mathbf{V}_\lambda &= \mathbf{I}_M - \mathbf{X}_\lambda(\mathbf{X}_\lambda^T\mathbf{X}_\lambda)^{-1}\mathbf{X}_\lambda^T,\\ \delta_\lambda^2 &= (\mathbf{1}^T\mathbf{V}_\lambda\mathbf{1})^{-1}\\ \tilde{m}_\lambda &= \delta_\lambda^2\mathbf{1}^T\mathbf{V}_\lambda\boldsymbol{y}_\lambda,\\ \nu &= M - 1 - n(\boldsymbol{u})\\ s_\lambda^2 &= \nu^{-1}\delta_\lambda^2\left[\boldsymbol{y}_\lambda^T\mathbf{V}_\lambda\boldsymbol{y}_\lambda - \delta_\lambda^2(\mathbf{1}^T\mathbf{V}_\lambda\boldsymbol{y}_\lambda)^2\right].\end{aligned} \tag{2.26}$$

En introduisant ces notations dans l'expression de la distribution *a posteriori* marginalisée (2.25), on peut déduire que la distribution *a posteriori* conditionnelle du paramètre σ_λ^2 est une loi inverse gamma définie par :

$$\sigma_\lambda^2|(m_\lambda, \boldsymbol{y}_\lambda, \boldsymbol{u}) \sim \mathcal{IG}\left(\frac{M-n(\boldsymbol{u})}{2}, \frac{1}{2\delta_\lambda^2}(\nu s_\lambda^2 + (m_\lambda - \tilde{m}_\lambda)^2)\right) \tag{2.27}$$

En calculant $\int_{\sigma_\lambda^2} p(m_\lambda, \sigma_\lambda^2|\boldsymbol{u}, \boldsymbol{y}_\lambda) d\sigma_\lambda^2$, *i.e.* en marginalisant en σ_λ^2, on obtient la distribution *a posteriori* conditionnelle du paramètre m_λ :

$$p(m_\lambda|\boldsymbol{y}_\lambda, \boldsymbol{u}) \propto \left(1 + \frac{1}{\nu}\left(\frac{m_\lambda - \tilde{m}_\lambda}{s_\lambda}\right)^2\right)^{-\frac{\nu+1}{2}}, \tag{2.28}$$

qui est une loi de Student avec ν degrés de liberté et où $\tilde{m}_\lambda$ est le paramètre de localisation et s_λ le paramètre d'échelle.

Comme la loi des paramètres m_λ et σ_λ^2 sont bien définies et que l'on sait générer des échantillons selon ces lois, on utilise un échantillonneur de Gibbs pour créer les chaînes de Markov $\{m_\lambda^{(k)}\}_k$ et $\{\sigma_\lambda^{2\,(k)}\}_k$ selon le schéma décrit dans l'encadré 2.1.

ENCADRÉ 2.1 – Echantillonnage de Gibbs

Initialisation : $m_\lambda^{(0)}$ et $\sigma_\lambda^{2\,(0)}$ fixés à la valeur estimée par méthode de σ-clipping pour chaque image à $1 \leqslant \lambda \leqslant \Lambda$ fixé.

Itération k **:** Pour tout $1 \leqslant \lambda \leqslant \Lambda$, connaissant $m_\lambda^{(k-1)}$ et $\sigma_\lambda^{2\,(k-1)}$:

1. Actualiser tous les paramètres qui ne varient pas lors de cette itération :
   - $\mathbf{X}^{(k)} = \mathbf{X}^{(k-1)}$ (*i.e.* $\boldsymbol{u}^{(k)} = \boldsymbol{u}^{(k-1)}$)
   - $\mathbf{V}_\lambda^{(k)} = \mathbf{V}_\lambda^{(k-1)}$
   - $\delta_\lambda^{2\,(k)} = \delta_\lambda^{2\,(k-1)}$
   - $\tilde{m}_\lambda^{(k)} = \tilde{m}_\lambda^{(k-1)}$
   - $s_\lambda^{2\,(k)} = s_\lambda^{2\,(k-1)}$
2. Générer $m_\lambda^{(k)} = s_\lambda\left(t^{(k)} + \tilde{m}_\lambda\right)$ où $t^{(k)}$ est tiré selon une loi de Student à ν degrés de liberté.
3. Générer $\sigma_\lambda^{2\,(k)}$ selon $\mathcal{IG}\left(\frac{M-n(\boldsymbol{u})}{2}, \frac{1}{2\delta_\lambda^{2\,(k)}}\left(\nu s_\lambda^{2\,(k)} + \left(m_\lambda^{(k)} - \tilde{m}_\lambda^{(k)}\right)^2\right)\right)$.
4. Evaluer $p(\boldsymbol{u}^{(k)}, \boldsymbol{m}^{(k)}, \boldsymbol{\sigma^2}^{(k)}|\mathbf{Y}) \propto \prod_{\lambda=1}^{\Lambda}\left\{p(m_\lambda^{(k)}, \sigma_\lambda^{2\,(k)}, \boldsymbol{u}^{(k)}|\boldsymbol{y}_\lambda)\right\}$.

Si la valeur de la densité *a posteriori* $p(\boldsymbol{u}^{(k)}, \boldsymbol{m}^{(k)}, \boldsymbol{\sigma^2}^{(k)}|\mathbf{Y})$ est plus élevée que la valeur maximale $p(\boldsymbol{u}^{(MAP)}, \boldsymbol{m}^{(MAP)}, \boldsymbol{\sigma^2}^{(MAP)}|\mathbf{Y})$ définie par :

$$p(\boldsymbol{u}^{(MAP)}, \boldsymbol{m}^{(MAP)}, \boldsymbol{\sigma^2}^{(MAP)}|\mathbf{Y}) = \max_{0 \leqslant i \leqslant k-1} p(\boldsymbol{u}^{(i)}, \boldsymbol{m}^{(i)}, \boldsymbol{\sigma^2}^{(i)}|\mathbf{Y})$$

alors on actualise l'estimateur MAP des différents paramètres inconnus :

- $\boldsymbol{u}^{(MAP)} = \boldsymbol{u}^{(k)}$
- $m_\lambda^{(MAP)} = m_\lambda^{(k)}$
- $\sigma_\lambda^{2\,(MAP)} = \sigma_\lambda^{2\,(k)}$

2.5.2 Echantillonnage de la configuration d'objet

La configuration d'objets ne peut pas être échantillonnée de la même façon que les paramètres du bruit. En effet, il n'est pas possible de générer directement un ensemble de configurations d'objets distribuées selon la loi *a posteriori* conditionnelle de la configuration $\boldsymbol{u}$. Puisque la dimension de l'espace des configurations varie (le nombre d'objets est a priori inconnu), nous avons fait le choix d'échantillonner la configuration à partir d'une méthode RJMCMC, en utilisant un échantillonneur de Metropolis-Hastings-Green (MHG). Cela nous permet de construire une suite de configurations d'objets sous la forme d'une chaîne de Markov ; chaque nouvelle configuration $\boldsymbol{v}$ est générée à partir de la précédente $\boldsymbol{u}$ à l'aide d'une loi instrumentale simple $q(\boldsymbol{u}, \boldsymbol{v}) = q(\boldsymbol{v}|\boldsymbol{u})$. Cette loi instrumentale q peut définir différents types de mouvements qui seront décrits dans les paragraphes suivants. Chaque mouvement qui modifie la dimension du problème doit être réversible, donc si l'on propose d'ajouter un objet, il faut aussi être en mesure de proposer la suppression de cet objet.

2.5.2.1 Mouvement de naissance-mort

Le mouvement de naissance-mort permet d'explorer un ensemble de configurations dont le nombre d'objets varie. Pour une configuration d'objets $\boldsymbol{u}$ donnée, notons $p_B(\boldsymbol{u})$ la probabilité de proposer une naissance et $p_D(\boldsymbol{u}) = 1 - p_B(\boldsymbol{u})$ est la probabilité de proposer une mort. En général $p_B(\boldsymbol{u}) = p_D(\boldsymbol{u}) = 1/2$. Dans le cas d'une mort l'objet supprimé $u_i \in \boldsymbol{u}$ est sélectionné avec une probabilité $p_S(u_i|\boldsymbol{u}) = \frac{1}{n(\boldsymbol{u})}$.

Dans le cas d'une naissance, notons $\boldsymbol{v} = \boldsymbol{u} \cup \{v\}$ la configuration proposée, avec v le nouvel objet proposé selon la mesure d'intensité $\nu(\cdot)$ du processus de référence et ses marques sont tirées aléatoirement suivant la densité décrite par l'*a priori* uniforme (2.19). Le ratio MHG s'écrit :

$$r(\boldsymbol{u}, \boldsymbol{v}) = \frac{p_D(\boldsymbol{v})}{p_B(\boldsymbol{u})} \frac{p(\boldsymbol{v}, \boldsymbol{m}, \boldsymbol{\sigma^2}|\mathbf{Y})}{p(\boldsymbol{u}, \boldsymbol{m}, \boldsymbol{\sigma^2}|\mathbf{Y})} p_S(v|\boldsymbol{v}). \tag{2.29}$$

Dans le cas d'une mort, un objet u_i, sélectionné avec la probabilité $p_S(u_i|\boldsymbol{u})$ est supprimé et la configuration proposée se note $\boldsymbol{v} = \boldsymbol{u} \backslash \{u_i\}$. Le ratio MHG correspondant s'écrit :

$$r(\boldsymbol{u}, \boldsymbol{v}) = \frac{p_B(\boldsymbol{v})}{p_D(\boldsymbol{u})} \frac{p(\boldsymbol{v}, \boldsymbol{m}, \boldsymbol{\sigma^2}|\mathbf{Y})}{p(\boldsymbol{u}, \boldsymbol{m}, \boldsymbol{\sigma^2}|\mathbf{Y})} \frac{1}{p_S(u_i|\boldsymbol{u})}. \tag{2.30}$$

Le mouvement de naissance ou de mort est ensuite accepté avec une probabilité $\alpha = \min(1, r(\boldsymbol{u}, \boldsymbol{v}))$.

Nous présentons ici l'implémentation du mouvement de type naissance-mort, le mouvement de naissance est détaillé dans l'encadré 2.2 et le mouvement de mort dans l'encadré 2.3. Nous avons mis en place une méthode de mise à jour récursive des différents termes qui apparaissent dans la densité *a posteriori*, équation (2.24). L'implémentation de cette mise à jour récursive pour les mouvements de naissance et de mort est détaillée dans l'annexe D. Elle repose sur la décomposition de Cholesky de la matrice de Gram $\mathbf{X}^T\mathbf{X} = CC^T$ et de $\left| \left(\mathbf{X}_\lambda^T \mathbf{X}_\lambda\right)^{-1} \right|$ qui apparaissent dans l'équation (2.25) où C est la matrice triangulaire inférieure de la décomposition de Cholesky. Cette stratégie de mise à jour récursive pour des mouvements de naissance-mort permet de réduire la complexité du calcul du ratio de MHG de $\mathcal{O}(n(\boldsymbol{u})^3)$ à $\mathcal{O}(n(\boldsymbol{u})^2)$, ce qui, en pratique, rend viable la méthode pour la détection de plusieurs centaines d'objets.

ENCADRÉ 2.2 – Mouvement de naissance

Si une naissance a été choisie avec une probabilité $p_B = 1/2$:

1. Sélectionner un point (p_0, q_0) dans la classe de pixels $\mathcal{C}_i$ avec la probabilité $\frac{p_{\mathcal{C}_i}}{|\mathcal{C}_i|}$ où $p_{\mathcal{C}_i}$ est la probabilité de choisir la classe $\mathcal{C}_i$ et $|\mathcal{C}_i|$ est le cardinal de la classe $\mathcal{C}_i$. Comme on se laisse la possibilité de proposer des positions de façon continue dans la grille de pixel, on ajoute une perturbation $(\Delta_p, \Delta_q) \sim \mathcal{U}([0,1[\times[0,1[)$ au point de coordonnées entières (p_0, q_0). Les coordonnées du nouvel objet v sont : $(p_v, q_v) = (p_0 + \Delta_p, q_0 + \Delta_q)$
2. Générer les marques de l'objet v :
   — $a_v \sim \mathcal{U}(r_{min}, r_{max})$
   — $b_v \sim \mathcal{U}(r_{min}, r_{max})$
   — $\alpha_v \sim \mathcal{U}(0, \frac{\pi}{2})$
   — $n_v \sim \mathcal{U}\{0.5, 1, 2\}$
3. Création de la nouvelle configuration $\boldsymbol{v} = \boldsymbol{u} \cup \{v\}$ que l'on traduit dans le langage matriciel par la matrice de configuration $\widetilde{\mathbf{X}} = \left[\mathbf{X}, \boldsymbol{x}_{n(\boldsymbol{v})}\right]$ où $\boldsymbol{x}_{n(\boldsymbol{v})}$ est le profil d'intensité vectorisé et normalisé de l'objet v.
4. Evaluation du ratio MHG après simplification :

$$
\begin{aligned}
r(\boldsymbol{u}, \boldsymbol{v}) &= q\left(\frac{\det \mathbf{X}^T\mathbf{X}}{\det \widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}}\right)^{\frac{\Lambda}{2}} \prod_{\lambda=1}^{\Lambda} \frac{\exp\left\{\frac{(\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda)^T\left(\widetilde{\mathbf{X}}(\widetilde{C}\widetilde{C}^T)^{-1}\widetilde{\mathbf{X}}^T\right)(\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda)}{2\sigma_\lambda^2}\right\}}{\exp\left\{\frac{(\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda)^T(\mathbf{X}(CC^T)^{-1}\mathbf{X}^T)(\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda)}{2\sigma_\lambda^2}\right\}} \times \sqrt{2\pi\sigma_\lambda^2} \\
&= q \times g^{-\Lambda} \times \prod_{\lambda=1}^{\Lambda} \exp\left\{\frac{a_{\lambda+1}^2 - 2m_\lambda a_1 a_{\lambda+1} + m_\lambda^2 a_1^2}{2\sigma_\lambda^2}\right\} \times \sqrt{2\pi\sigma_\lambda^2}
\end{aligned}
\tag{2.31}
$$

   où
   — $\boldsymbol{a} = \frac{1}{g}\left(x_{n(\boldsymbol{u})+1} - \boldsymbol{v_p}^T C^{-1}\mathbf{X}^T\right)[\mathbf{1}, \boldsymbol{y}_1, ..., \boldsymbol{y}_\Lambda]$ est un vecteur ligne de taille $\Lambda + 1$, et a_i est sa $i^{\text{ème}}$ composante.
   — $v = \boldsymbol{x}_{n(\boldsymbol{u})+1}^T \boldsymbol{x}_{n(\boldsymbol{u})+1}$,
   — $\boldsymbol{v_p} = C^{-1}\mathbf{X}^T\boldsymbol{x}_{n(\boldsymbol{u})+1}$,
   — $g = \sqrt{v - \boldsymbol{v_p}^T\boldsymbol{v_p}}$,
   Les calculs sont détaillés en annexe dans le paragraphe D.1
5. Phase d'acceptation-rejet et mise à jour de la nouvelle configuration.

ENCADRÉ 2.3 – Mouvement de mort

Si une mort a été choisie avec une probabilité $p_D = 1/2$, si la configuration courante $\boldsymbol{u}$ est vide alors ne rien faire, sinon :

1. Sélectionner un objet de la configuration courante $u_i \in \boldsymbol{u}$ avec la probabilité $p_S(u_i|\boldsymbol{u})$.
2. Création de la nouvelle configuration $\boldsymbol{v} = \boldsymbol{u}\backslash\{u_i\}$ que l'on traduit dans le langage matriciel par la matrice de configuration $\widetilde{\mathbf{X}} = \mathbf{X}_{\backslash u_i}$ où $\mathbf{X}_{\backslash u_i}$ est la matrice $\mathbf{X}$ privée de la colonne correspondant à l'objet sélectionné u_i.
3. Evaluation du ratio MHG après simplification :

$$r(\boldsymbol{u}, \boldsymbol{v}) = g^{\Lambda} \times \frac{1}{q} \times \prod_{\lambda=1}^{\Lambda} \exp\left\{\frac{-a_{\lambda+1}^2 + 2m_\lambda a_1 a_{\lambda+1} - m_\lambda^2 a_1^2}{2\sigma_\lambda^2}\right\} \times \frac{1}{\sqrt{2\pi\sigma_\lambda^2}} \tag{2.32}$$

   Les calculs permettant d'écrire cette expression simplifiée sont détaillés en annexe dans le paragraphe D.2.

4. Phase d'acceptation-rejet et mise à jour de la nouvelle configuration.

2.5.2.2 Mouvements simples sur un objet de la configuration courante

Dans le cas de mouvements simples tels que la rotation, la translation ou la modification de la forme d'un objet $u_i \in \boldsymbol{u}$, la configuration ne change pas de dimension. L'algorithme d'échantillonnage correspond donc à l'algorithme de Metropolis-Hastings (MH) classique (voir le principe dans le paragraphe B.2.2). Le principe de ces mouvements est d'améliorer la modélisation des données par les objets en leur faisant subir des petites perturbations telles que la figure 2.3 les illustre. La procédure est assez similaire à celles mises en oeuvre pour les mouvements de naissance et de mort. Un objet est sélectionné avec la probabilité $p_S(u_i|\boldsymbol{u})$ et un type de modification est sélectionné avec la probabilité correspondante :

— p_r est la probabilité de sélectionner une rotation.
— p_t est la probabilité de sélectionner une translation.
— p_m est la probabilité de sélectionner une modification de forme.

L'indice Sersic n du profil d'intensité de l'objet sera échantillonné en même temps que la modification de forme.

Les marques du nouvel objet v sont modélisées à partir des marques de l'objet u_i sélectionné qui subissent des petites perturbations par marche aléatoire. Cette dépendance se traduit par la densité $q(v|u_i)$. Le ratio MH correspondant s'écrit :

$$r(\boldsymbol{u}, \boldsymbol{v}) = \frac{p_S(v|\boldsymbol{v})}{p_S(u_i|\boldsymbol{u})} \frac{p(\boldsymbol{v}, \boldsymbol{m}, \boldsymbol{\sigma^2}|\mathbf{Y})}{p(\boldsymbol{u}, \boldsymbol{m}, \boldsymbol{\sigma^2}|\mathbf{Y})} \frac{q(u_i|v)}{q(v|u_i)}, \tag{2.33}$$

Ce type de mouvement peut se voir comme la mort de l'objet u_i et la naissance de sa version modifiée v. Il est donc possible de calculer le ratio MHG comme une combinaison de ratios MHG de naissance puis de mort en prenant soin d'ajouter la densité $q(v|u_i)$ qui traduit la dépendance entre les objets u_i et v qui est réalisée ici par une marche aléatoire sur les marques des objets (position, largeurs à mi-hauteur, orientation). L'implémentation est basée sur cette stratégie, ce qui permet l'utilisation de la mise à jour récursive des matrices intervenant dans le calcul du ratio.

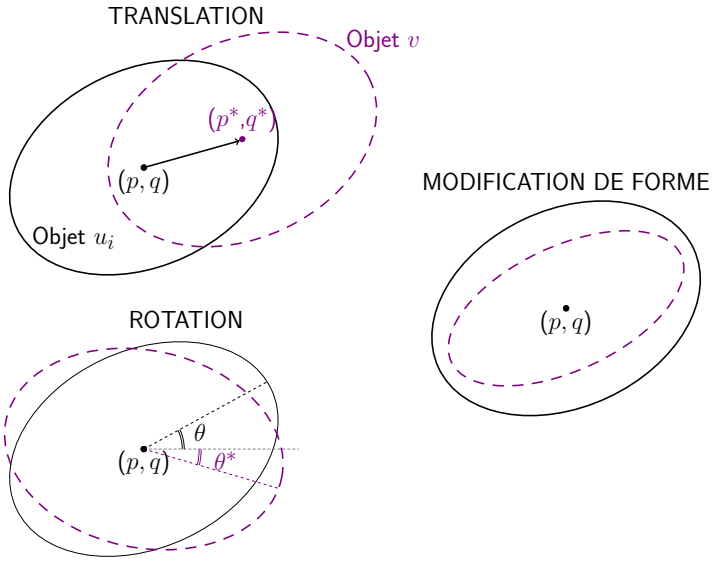


FIGURE 2.3 – Illustration des différents mouvements possibles sur un objet u_i (trait plein noir) sélectionné dans la configuration courante $\boldsymbol{u}$. Cet objet sera remplacé par l'objet v (trait pointillé violet) pour former la nouvelle configuration $\boldsymbol{v} = \boldsymbol{u} \backslash u_i \cup v$

Modification d'un objet de la configuration courante

1. Sélectionner un objet de la configuration courante $u_i \in \boldsymbol{u}$ avec la probabilité $p_S(u_i|\boldsymbol{u})$.
2. Sélectionner une transformation selon la probabilité p_r, p_t ou p_m.
3. Générer l'objet v selon $q(v|u_i)$.
4. Mort de l'objet u_i et naissance de l'objet v.
5. Phase d'acceptation-rejet et mise à jour de la nouvelle configuration.

2.6 Structure de l'algorithme de détection

L'échantillonnage de la configuration d'objets par méthode RJMCMC peut s'avérer très coûteux en temps de calcul de par la nature itérative de l'algorithme et de par la dimension des données. Nous allons présenter dans ce paragraphe les différentes stratégies mises en oeuvre afin de réduire ce temps de calcul.

2.6.1 Détection des objets les plus brillants sur l'image blanche

La détection des objets à spectre continu et des objets dont le spectre ne contient que des raies d'émissions peut se faire en deux temps :

— détection sur l'image blanche pour les objets à spectre continu (ou contenant une raie d'émission d'intensité suffisamment forte pour subsister après moyennage des Λ longueurs d'onde,

— détection sur le cube complet pour les objets contenant une ou plusieurs raies d'émission de plus faible intensité et dont la position dans le spectre est inconnue.

Bien sûr les objets à spectre continu peuvent être détectés par notre algorithme RJMCMC sur le cube complet, mais les détecter sur une simple image permet un gain de temps de calcul

considérable. De plus il est possible d'utiliser des méthodes rapides et largement adoptées par la communauté astrophysique (voir chapitre 1) ; la seule contrainte que nous imposons est que la méthode utilisée doit produire un profil spatial d'intensité en deux dimensions afin que nous puissions construire la matrice $\mathbf{X}$ qui contient aussi bien les profils des objets à spectre continu que ceux des objets à raies d'émission. Il existe des méthodes plus rapides dans la littérature pour la détection de sources sur une image seule. On peut par exemple utiliser SExtractor (Bertin and Arnouts [1996]) pour détecter les objets présents sur l'image blanche. Cette méthode retourne la localisation des objets détectés. SExtractor fournit également un grand nombre de paramètres astrophysiques pour décrire les objets détectés. Nous ne décrirons pas toutes les possibilités offertes par SExtractor, on peut les retrouver dans le guide de l'utilisateur fourni par le concepteur de la méthode, Bertin, et dans le guide rédigé par Holwerda. Le logiciel SExtractor ne fournit pas d'estimation du profil spatial d'intensité, en revanche il peut, par exemple, fournir les paramètres de l'ellipse modélisant l'objet ou encore la largeur à mi-hauteur du profil. En imposant un profil spatial d'intensité de la forme d'une gaussienne en deux dimensions et en se servant du support elliptique renvoyé par SExtractor, nous pourrions créer la matrice $\mathbf{X}$ pour les objets détectés sur l'image blanche. Nous avons seulement besoin du profil puisqu'il est ensuite normalisé et que l'intensité correspondante est estimée par la suite. Dans cette thèse, nous n'avons pas implémenté cette solution qui nécessite des connaissances approfondies en astrophysique pour transposer les paramètres de sortie de SExtractor en paramètres exploitable par notre méthode.

Il est tout à fait possible d'utiliser la méthode de détection basée sur un processus ponctuel marqué et un modèle bayésien pour effectuer la détection sur l'image blanche. Soit $\boldsymbol{y}_{white} \in \mathbb{R}^M$ l'image blanche sous forme vectorisée, le modèle d'observation peut s'écrire :

$$\boldsymbol{y}_{white} = \mathbf{X}_{white}\boldsymbol{w}_{white} + \mathbf{1}m_{white} + \boldsymbol{\epsilon}_{Bg,white}$$

où $\boldsymbol{w}_{white}$ est le vecteur contenant les intensités des objets détectés sur l'image blanche, m_{white} et σ^2_{white} sont respectivement la moyenne et la variance du bruit de fond présent sur l'image blanche. Ce bruit est gaussien puisque :

$$\boldsymbol{y}_{white} = \frac{1}{\Lambda}\sum_{\lambda=1}^{\Lambda} \boldsymbol{y}_\lambda$$

et donc :

$$\boldsymbol{\epsilon}_{Bg,white} = \frac{1}{\Lambda}\sum_{\lambda=1}^{\Lambda} \boldsymbol{\epsilon}_{Bg,\lambda}.$$

Le vector $\boldsymbol{\epsilon}_{Bg,white}$ est un vecteur aléatoire gaussien puisqu'il est la somme de vecteurs aléatoires gaussiens indépendants. Les *a priori* sur les paramètres du bruit, sur les objets et les relations inter-objets et sur leurs intensités utilisés dans le cas de l'image blanche sont les mêmes que ceux utilisés sur une image correspondant à une longueur d'onde λ donnée du cube. Ce qui nous conduit à écrire une densité *a posteriori* à échantillonner très similaire à la densité associée au modèle sur le cube complet :

$$\begin{aligned} p(\boldsymbol{u}, m_{white}, \sigma^2_{white}|\boldsymbol{y}_{white}) &\propto\ p(m_{white}, \sigma^2_{white}|\boldsymbol{u}, \boldsymbol{y}_{white})p(\boldsymbol{u}) \\ &\propto \left(\frac{1}{2\pi\sigma^2_{white}}\right)^{\frac{M-n(\boldsymbol{u})}{2}} e^{-\frac{(\boldsymbol{y}_{white}-\mathbf{1}m_{white})^T\mathbf{V}_{white}(\boldsymbol{y}_{white}-\mathbf{1}m_{white})}{2\sigma^2_{white}}} \\ &\quad \times \left|\mathbf{X}^T_{white}\mathbf{X}_{white}\right|^{-\frac{1}{2}} \times \frac{1}{\sigma^2_{white}}\mathbb{1}_{]0,+\infty[}(\sigma^2_{white}) \\ &\quad \times \Gamma\left(n(\boldsymbol{u})+1\right) \times q^{n(\boldsymbol{u})+1}h(\boldsymbol{u}). \end{aligned} \tag{2.34}$$

avec $\mathbf{V}_{white} = \mathbf{I}_M - \mathbf{X}_{white}(\mathbf{X}_{white}^T\mathbf{X}_{white})^{-1}\mathbf{X}_{white}^T$. Il suffit alors d'utiliser l'algorithme d'échantillonnage sur l'image blanche. Cette stratégie possède l'avantage de fournir en sortie un catalogue d'objets parfaitement formatés pour initialiser la configuration d'objets lors de la détection sur le cube complet. En effet, la matrice $\mathbf{X}_{white}$ a déjà le bon format puisque chacune de ses colonnes contient la réponse normalisée ℓ_2 de l'instrument à chacune des sources visibles et détectées sur l'image blanche. L'inconvénient majeur de cette approche réside dans son temps d'exécution.

2.6.2 Algorithme de détection

La structure globale de l'algorithme est illustrée par la figure 2.4. Elle est finalement composée de deux chaînes parallèles de prétraitement qui permettent d'initialiser le processus de détection par échantillonnage de la configuration d'objets avec la méthode RJMCMC.

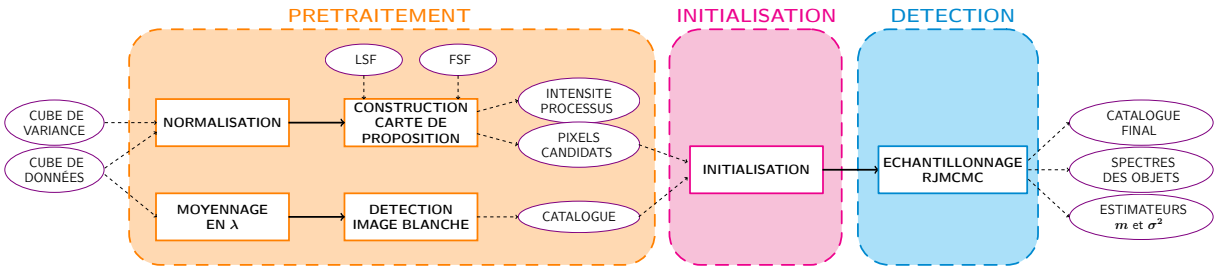


FIGURE 2.4 – Description de la structure de la chaîne de traitement des données MUSE jusqu'à la production du catalogue d'objets

Ces deux chaînes de prétraitement permettent de réduire considérablement le temps de calcul de l'algorithme principal. Les résultats intermédiaires produits lors du prétraitement sont :

- une carte de proposition, *i.e.* l'ensemble des pixels proposés pour être un centre d'objet,
- la fonction d'intensité du processus ponctuel,
- un catalogue d'objets contenant une composante continue ou une raie d'émission avec une intensité suffisamment forte pour lui permettre de survivre au moyennage des données en longueurs d'onde.

La phase d'initialisation consiste à fixer dans la configuration un certain nombre d'objets qui ne seront plus modifiés par la suite. On considère que ces objets détectés sur l'image blanche sont les meilleurs estimateurs possibles, au sens du maximum *a posteriori* s'ils ont été détectés à l'aide de la méthode de détection basée sur un processus ponctuel marqué et le modèle bayésien décrit par l'équation (2.34), ou au sens d'un autre critère selon la méthode utilisée. Nous verrons par la suite (chapitre 3 et chapitre 4) que la fonction d'intensité du processus est élaborée dans le but de favoriser les objets au niveau des galaxies ne contenant pas de composante continue, mais seulement une voire plusieurs raies. Autoriser la modification des objets détectés sur l'image blanche pourrait alors entraîner la modification, la suppression ou la surdétection de certains d'entre eux, ce qui n'est pas souhaitable.

Les résultats produits par l'étape de détection sur le cube complet et plus généralement par l'algorithme sont :

- un catalogue d'objets contenant leurs caractéristiques,
- l'estimation des intensités des objets au sens du maximum *a posteriori*, sachant la configuration d'objets détectés, à chaque longueur d'onde λ : $\boldsymbol{w}_\lambda = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}_\lambda$ où $\boldsymbol{w}_\lambda$ est une colonne de la matrice définie par l'équation (2.12). Les spectres des objets sont les lignes de cette matrice,
- l'estimation au sens du *maximum a posteriori* des paramètres du bruit (moyenne et variance).

2.6.3 Parallélisation de l'échantillonnage

Nous avons vu dans le paragraphe 2.5.2 que lorsque la configuration d'objets est modifiée, le calcul du critère d'acceptation-rejet et la mise à jour de la valeur de la densité *a posteriori* pouvaient être effectués de façon récursive en ne calculant que les coefficients de la matrice $\mathbf{X}^T\mathbf{X}$ et de $(\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda)^T(\mathbf{X}^T\mathbf{X})^{-1}(\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda)$ impactés par le changement, pour tout $1 \leqslant \lambda \leqslant \Lambda$. En revanche lorsque ce sont les paramètres de moyenne et de variance du bruit qui sont mis à jour alors il faut recalculer l'ensemble les différentes matrices et valeurs de façon directe et générer les paramètres m_λ et σ^2_λ à chaque longueur d'onde. Du fait de l'hypothèse d'indépendance du bruit longueur d'onde par longueur d'onde, on peut paralléliser les Λ longueurs d'onde sur lesquelles l'échantillonnage de Gibbs et l'évaluation de $(\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda)^T(\mathbf{X}^T\mathbf{X})^{-1}(\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda)$ doivent être réalisés.

2.7 Discussion sur la méthode

Nous avons décrit le processus d'échantillonnage de la configuration d'objets et des paramètres du bruit de fond dans les paragraphes précédentes. Nous allons maintenant discuter certains points clés du choix de la modélisation par processus ponctuel marqué et de l'échantillonnage de ce processus par méthode RJMCMC.

2.7.1 Critère d'arrêt

Si les méthodes RJMCMC convergent d'un point de vue théorique vers la loi cible, lors de l'implémentation d'une telle méthode, il est difficile de traduire cette convergence à l'aide d'une condition d'arrêt bien définie. Comment décider alors à quel moment stopper l'échantillonnage ? Il faut que le nombre d'itérations, c'est-à-dire le nombre d'éléments des chaînes de Markov produites par l'algorithme RJMCMC, soit suffisamment grand pour garantir la convergence de la distribution de ces échantillons vers leur loi cible. Une des solutions consiste à fixer un nombre maximal d'itérations très grand devant la dimension du problème et le nombre de paramètres à échantillonner. Cependant, les algorithmes de type RJMCMC étant itératifs, et le temps de calcul d'une itération étant relativement important, il n'est peut être pas nécessaire de générer plusieurs dizaines, voire centaines, de milliers d'itérations avant d'obtenir un résultat tout à fait satisfaisant. Il faut donc établir un critère d'arrêt de l'échantillonnage de la chaîne de Markov. D'un point de vue de l'implémentation, nous sommes aussi obligés de fixer un nombre maximum d'itérations afin de gérer l'allocation de mémoire.

Observons notre problème : nous cherchons à échantillonner la distribution *a posteriori* de tous les paramètres du modèle, et en particulier à estimer la configuration d'objets. Nous souhaitons donc que la chaîne de Markov explore l'ensemble des configurations d'objets les plus probables afin d'en extraire celle qui maximise la densité *a posteriori* définie par (2.25). La configuration estimée au sens du maximum *a posteriori* (MAP) et la densité *a posteriori* évaluée pour cette configuration sont mises à jour à chaque itération. Si la nouvelle configuration proposée à l'itération k maximise la densité *a posteriori* par rapport aux $k-1$ premières valeurs échantillonnées, alors cette configuration devient l'estimation MAP. Pour décider quand arrêter l'échantillonnage RJMCMC, il faut que la configuration MAP atteigne un nombre stable d'objets. Nous fixons donc un critère d'arrêt sur la stabilité de la statistique du nombre d'objets de l'estimation MAP de la configuration d'objets : si aucune mort ni aucune naissance ne sont acceptées durant K itérations, l'algorithme d'échantillonnage s'arrête. On définit K par le critère heuristique suivant :

$$K = 3 \times \#\{\text{pixels de la carte de proposition}\}\frac{1}{\text{Proba(Naissance)}}, \tag{2.35}$$

où #{pixels de la carte de proposition} est le nombre de pixels de la carte de proposition et Proba(Naissance) est la proportion de mouvements de naissance proposés par rapport aux autres mouvements de l'algorithme d'échantillonnage. On s'assure ainsi qu'en moyenne on a proposé trois fois [1] tous les pixels de la carte de proposition et qu'aucune naissance ni aucune mort d'objet n'ont été acceptées avant de stopper l'algorithme. Seuls des mouvements de translation, rotation ou modification de la taille des axes et de l'indice Sersic ont pu être acceptés durant ces K dernières itérations (et bien sûr les paramètres du bruit ont aussi été échantillonnés durant ces itérations). On ne peut fixer de critère d'arrêt sur ces autres mouvements car du fait des erreurs de modèle, de l'échantillonnage du profil Sersic, ou encore des erreurs de modélisation de la PSF de l'instrument, des mouvements modifiant la forme des objets sont toujours possibles.

Finalement, malgré le caractère stochastique de la méthode, en effectuant plusieurs fois le processus de détection décrit sur la figure 2.4 avec les mêmes paramètres d'entrées, le nombre d'objets obtenus dans les différentes configurations estimées est stable [2]. La plus grande variation provient de la détection sur l'image blanche pour laquelle l'influence du terme d'attache aux données est moindre comparée à la détection sur le cube complet et donc la loi du processus ponctuel est moins concentrée autour de sa valeur la plus probable. La configuration d'objets estimée sur l'image blanche est un peu moins stable que la configuration détectée sur le cube complet. Alors qu'il n'y a pas ou très peu de mouvements de mort acceptés lors de la détection sur le cube complet, sur l'image blanche les objets sont proposés, acceptés et supprimés plusieurs fois avant de rester définitivement dans la configuration estimée au sens du MAP. Il faut donc prendre garde à choisir un nombre maximum d'itérations suffisamment grand afin que ce nombre ne soit pas le facteur limitant de la détection sur l'image blanche à l'aide l'algorithme d'échantillonnage RJMCMC. Nous savons que les étoiles et les galaxies les plus brillantes (et donc relativement proches comparées aux galaxies jeunes que nous cherchons) détectables sur l'image blanche n'ont pas un support spatial et un profil spatial d'intensité réguliers, le modèle Sersic elliptique n'est donc pas forcément le plus approprié pour modéliser les sources visibles sur l'image blanche. Dans les implémentations futures de l'algorithme, nous pouvons envisager de nous servir de méthodes de détection développées spécifiquement pour des images telles que SExtractor ou GIM2D pour effectuer la détection sur l'image blanche. Les sources détectées seront sans doute mieux modélisées et mieux localisées, en revanche un grand nombre de galaxies de très faible intensité ne seront pas détectées par ces méthodes. Ces méthodes ne devront être utilisées que comme une étape de prétraitement.

2.7.2 Influence du modèle Sersic elliptique sur les erreurs d'estimation

Nous avons choisi de représenter les galaxies par une configuration d'objets simples qui serait une réalisation d'un processus ponctuel marqué. En adoptant un modèle simplifié pour les galaxies (support elliptique et profil de Sersic pour la décroissance d'intensité), nous assumons faire des erreurs de modélisation, notamment au niveau des galaxies les plus brillantes et les plus étendues spatialement. Afin de caractériser ces erreurs, nous allons travailler à partir de galaxies simulées par les astrophysiciens du consortium MUSE et mesurer les erreurs induites par le modèle Sersic elliptique. Ces galaxies sont le résultat de simulations numériques cherchant à reproduire des phénomènes physiques observés. Parmi les 18 galaxies simulées dans le cube DryRun, la galaxie

1. Ce nombre moyen a été fixé de manière empirique en observant que l'algorithme de détection appliqué à des cubes de données synthétiques et réels voit le nombre d'objets détectés varier rapidement tant que la phase de convergence n'est pas atteinte. Un intervalle de K itérations défini par l'équation (2.35) sur lequel le nombre d'objets de la configuration estimée au sens du maximum *a posteriori* reste stable nous assure de la convergence de l'algorithme.

2. Le nombre objet de la configuration estimée au sens du maximum *a posteriori* peut varier de quelques ($\leqslant 10$) objets d'une chaine RJMCMC à l'autre pour un nombre moyen d'objets supérieur à trois cents dans un cube de données MUSE de taille standard.

ID#5, représentée sur la figure 2.5, est un bon candidat pour estimer les erreurs de modélisation. En effet, elle présente un profil d'intensité non régulier du fait de la présence de deux maxima locaux sur le profil d'intensité spatial dus au champ de vitesse simulé lors de la création de cette galaxie. De plus sa forme n'est pas parfaitement elliptique comme on peut le voir sur la figure 2.5.

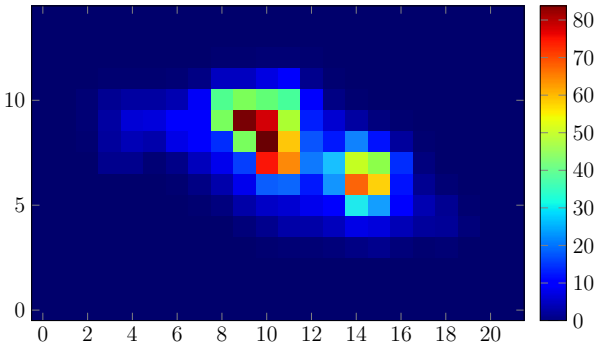


FIGURE 2.5 – Profil d'intensité de la galaxie ID#5 du cube DryRun en l'absence de bruit.

Plusieurs erreurs de modélisation peuvent apparaître sur une telle galaxie :

Erreur de sous-ajustement : la galaxie possède un profil d'intensité spatial avec deux maxima locaux, elle est détectée par un seul objet, le profil Sersic n'est alors pas adapté à la modélisation de l'intensité.

Erreur de sur-ajustement : la galaxie possède un profil d'intensité spatial avec deux maxima locaux, elle est détectée par deux petits objets indépendants qui modélisent beaucoup mieux l'intensité. Mais le nombre d'objets détectés ne correspond pas au nombre de galaxies observées.

Dans cette section, nous allons étudier l'influence de ces erreurs à l'aide du critère d'erreur quadratique moyenne :

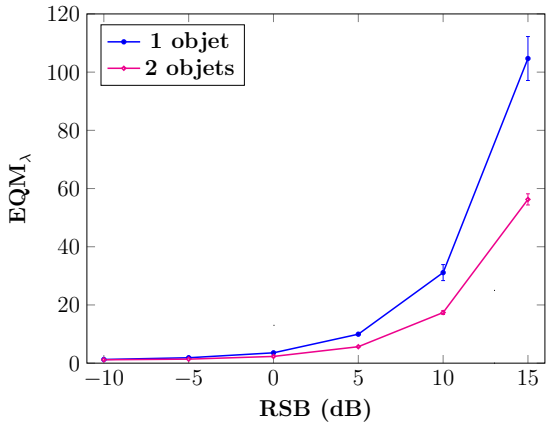
$$\mathrm{EQM}_\lambda = \mathrm{E}\left[(\boldsymbol{y}_\lambda - \mathbf{X}\boldsymbol{w}_\lambda)^2\right] \tag{2.36}$$

et de la proportion d'énergie de la galaxie modélisée par le ou les objets :

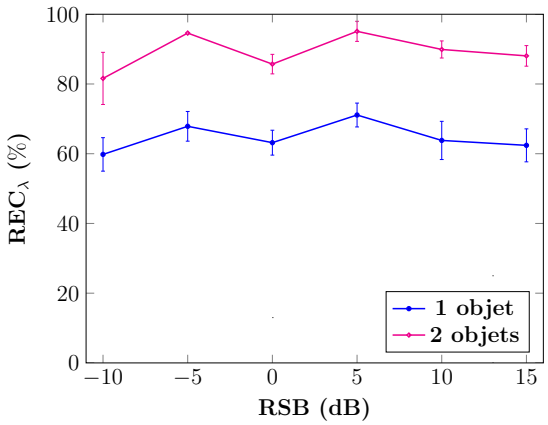
$$\mathrm{REC}_\lambda = \frac{\left\|\mathbf{X}\boldsymbol{w}_\lambda\right\|_2^2}{\left\|\boldsymbol{y}_\lambda - \epsilon_{Bg,\lambda}\right\|_2^2} \tag{2.37}$$

Pour simplifier l'analyse, nous allons considérer une seule image à λ fixé contenant une seule galaxie, cette image correspond à la longueur d'onde de la valeur maximum du spectre de la galaxie.

L'algorithme de détection est utilisé sur un cube contenant la galaxie ID#5 seule et un bruit gaussien centré réduit indépendant spatialement. Cent détections indépendantes sont effectuées sur ce cube. Les résidus de modélisation sont calculés à partir des cent configurations extraites (cent chaînes différentes) au sens du maximum *a posteriori* sur l'image correspondant à la longueur d'onde λ du maximum du spectre de la galaxie ID#5 (λ est connue, on dispose du spectre de l'objet). On effectue ces simulations pour différents niveaux de rapport signal à bruit (RSB définit selon la définition donnée par l'équation (1.10)) en modifiant l'intensité de la galaxie et en maintenant la moyenne du bruit à $m = 0$ et $\sigma^2 = 1$. Le prétraitement destiné à construire la carte de proposition fournit deux positions possibles pour le centre des objets proposés. Les configuration d'objets estimées au sens du MAP mettent en évidence tantôt des erreurs de sous-ajustement tantôt des erreurs de sur-ajustement. Les performances sont mesurées à l'aide des deux critères



(a) EQM$_\lambda$ moyenne obtenue pour différents RSB.



(b) REC$_\lambda$ moyen obtenu pour différents RSB.

FIGURE 2.6 – Performance de la détection dans le cas où la configuration estimée ne contient qu'un seul objet (courbes bleues) et dans le cas où la galaxie a été modélisée par deux objets (courbes magenta). Les barres d'erreurs modélisent l'écart-type des valeurs d'EQM$_\lambda$ et de REC$_\lambda$ obtenues sur les différentes simulations.

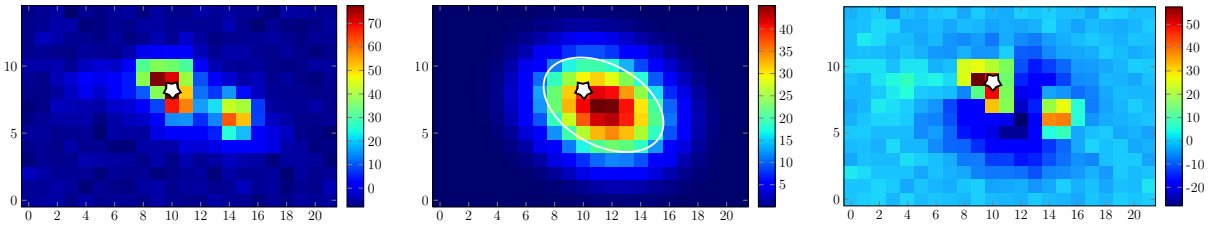
décrits par les équations (2.36) et (2.37) et sont représentés sur la figure 2.6. Plusieurs constats peuvent être dressés à partir de ces résultats :

1. A fort rapport signal à bruit, les erreurs de modélisation de type sous-ajustement sont plus gênantes que les erreurs de type sur-ajustement en ce qui concerne les deux critères d'évaluations (EQM$_\lambda$ et REC$_\lambda$). D'un point de vue traitement du signal, les résidus très forts au niveau des maxima locaux dans le profil d'intensité sont considérés par l'algorithme comme des sources à modéliser. Ces résidus peuvent produire un effet d'aveuglement de l'algorithme qui ne sera pas capable de détecter des galaxies de faible intensité proche de ces résidus.
2. A faible rapport signal à bruit, l'erreur quadratique moyenne est similaire pour les erreurs de type sous-ajustement et de sur-ajustement. En revanche, détecter une source comme la galaxie ID#5 par deux objets permet de mieux expliquer les données, notamment en terme d'énergie modélisée.
3. De manière générale, modéliser la galaxie ID#5 par deux objets centrés sur les maxima locaux du profil d'intensité permet d'expliquer au mieux la contribution de la galaxie dans l'image. Les configurations avec deux objets modélisent en moyenne 20% d'énergie en plus que les configurations avec un seul objet, et ce quel que soit le rapport signal à bruit considéré.

2.7.2.1 Mise en évidence des erreurs de modélisation de type sous-ajustement

La figure 2.7 représente les résultats obtenus pour l'une des cent détections effectuées à un niveau de bruit RSB = 15dB. Lorsque la détection estimée au sens du MAP ne comporte qu'un seul objet, le centre proposé correspond au maximum local de plus forte intensité. Il est modélisé par l'étoile blanche sur les figures 2.7a, 2.7b, 2.7c et 2.8. Ce centre correspond au pixel de plus forte intensité dans le profil de la galaxie. Bien que l'objet ait été proposé sur ce pixel, l'algorithme converge systématiquement vers une position moyenne entre les deux maxima locaux, modélisée par l'étoile rouge sur la figure 2.8. Il s'agit du meilleur compromis possible du point de vue de la

minimisation des erreurs de modélisation par un unique profil Sersic. Le déplacement de l'objet durant l'échantillonnage de la configuration est modélisé sur la figure 2.8.



(a) Image bruitée (b) Estimation (c) Erreur de reconstruction

Figure 2.7 – Estimation des erreurs de modélisation sur la galaxie simulée ID#5 du cube DryRun en présence d'un bruit gaussien centré réduit avec un rapport signal à bruit de 15dB. L'étoile blanche symbolise le centre proposé grâce à la carte de proposition lors de la naissance de l'objet.

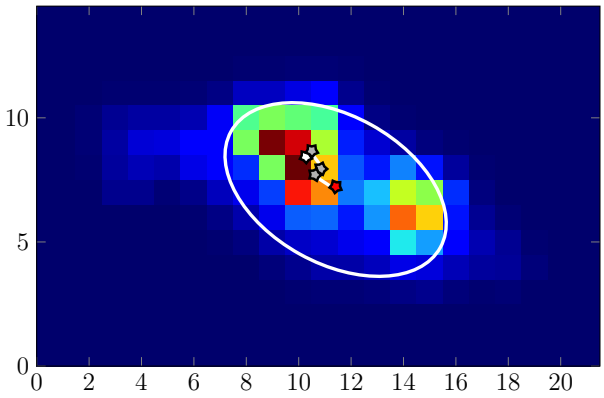
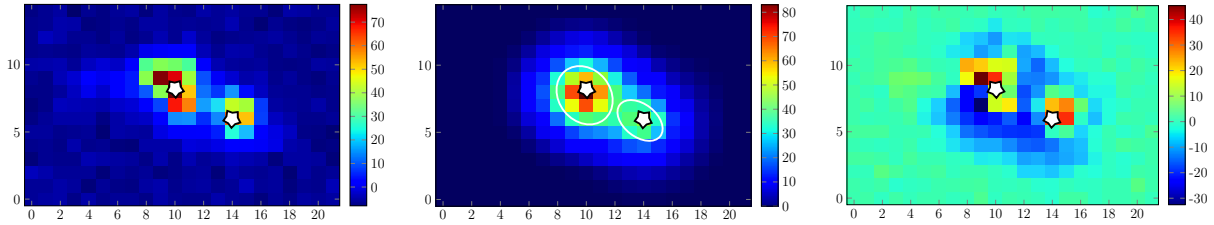


Figure 2.8 – Déplacement du centre de l'objet qui modélise la galaxie ID#5. L'ellipse blanche représente le support elliptique de l'objet proposé avant convolution avec la FSF moyenne. L'étoile blanche correspond à la position proposée lors de la naissance de l'objet, les étoiles grises marquent les positions intermédiaires et l'étoile rouge symbolise la position finale de l'objet.

2.7.2.2 Mise en évidence des erreurs de modélisation de type sur-ajustement

Dans le cas où la configuration estimée au sens du MAP contient deux objets, leur centre coïncide avec les deux maxima locaux du profil d'intensité spatial. Cela permet de mieux modéliser le profil complexe de la galaxie ID#5 du cube DryRun. Il reste cependant des résidus autour des deux objets qui sont dûs notamment à l'utilisation de profils Sersic qui ne modélisent pas parfaitement la décroissance d'intensité dans le cas de cette galaxie. Pour un fort rapport signal à bruit (ici 15dB), les amplitudes maximale et minimale des résidus générés par les erreurs de type sur-ajustement sont du même ordre de grandeur que pour les erreurs de type sous-ajustement. Cependant le cas présenté sur la figure 2.9 montre que les résidus sont moins étendus que dans le cas présenté sur la figure 2.7, ce qui explique que l'erreur quadratique moyenne soit plus faible dans le cas des erreurs de sur-ajustement. Ces conclusions sont, en moyenne, vérifiées sur les différentes détections effectuées sur cette source. Des conclusions inversées pourraient être tirées sur une source avec un profil spatial d'intensité atypique différent (par exemple une source présentant un seul maximum local mais une forme assez diffuse mais non elliptique).



(a) Image bruitée (b) Estimation (c) Erreur de reconstruction

FIGURE 2.9 – Estimation des erreurs de modélisation sur la galaxie simulée ID#5 du cube DryRun en présence d'un bruit gaussien centré réduit avec un rapport signal à bruit de 15dB. Les étoiles blanches symbolisent les centres proposés grâce à la carte de proposition lors des naissances des deux objets. Les ellipses blanches (b) modélisent le support elliptique des objets avant convolution par la FSF.

2.7.2.3 Bilan sur les erreurs de modélisation

D'un point de vue de la modélisation, les erreurs de type sur-ajustement sont moins gênantes que les erreurs de type sous-ajustement : elles permettent de modéliser plus précisément l'énergie apportée par une galaxie avec un profil d'intensité atypique, et favorisent ensuite la détection d'autres galaxies dans le voisinage, en particulier d'autres galaxies avec une intensité plus faible.

D'un point de vue astrophysique, nous comprenons bien que les erreurs de type sur-ajustement entraînent des erreurs d'interprétation en terme de nombre d'objets et de modélisation des paramètres physiques de la galaxie. De plus ce type d'erreur doit être limité puisque la validation des objets se fait en post-traitement, manuellement, en visitant chaque objet de la configuration estimée au sens du MAP. Notons que si le profil spatial d'intensité de la galaxie comporte plusieurs maxima locaux, le terme de pénalisation $h_2(\boldsymbol{u})$ introduit dans l'équation (2.21) est censé limiter les détections multiples, notamment lorsque les maxima locaux sont proches et que le centre de la galaxie concentre une grande proportion de son énergie.

Les résultats de détection montrés sur les figures 2.7 et 2.9 soulignent la nécessité de modéliser au mieux les galaxies de fort rapport signal à bruit afin de minimiser les résidus dont l'amplitude maximale peut être plus importante que l'amplitude des galaxies de faible intensité présentes dans le voisinage.

2.8 Bilan

Modélisation des galaxies

Une galaxie est modélisée par un profil d'intensité spatial Sersic, défini dans un repère elliptique et limité par un support elliptique. Des contraintes sont introduites dans la densité du processus ponctuel marqué afin d'interdire les recouvrements entre les objets trop importants.
La modélisation du profil d'intensité des galaxies par un profil Sersic à support elliptique c'est pas parfaite, en particulier pour les objets brillants et spatialement étendu, mais permet de modéliser simplement les galaxies lointaines.

Modèle d'observation

Le modèle d'observation utilisé pour les données MUSE repose sur quelques hypothèses :
- la configuration de galaxies est identique à toutes les longueurs d'onde, c'est l'intensité des galaxies qui gère la présence ou l'absence d'émission à une longueur d'onde donnée,
- le bruit est supposé gaussien i.i.d. spatialement et spectralement, seules les contributions des galaxies sont composées par la PSF de l'instrument.

Echantillonnage et estimation de la configuration d'objets

Afin d'échantillonner la configuration de galaxies, un algorithme itératif de type RJMCMC est utilisé. Les modifications de la configuration d'objets permettent de construire une chaîne de Markov de configurations, *i.e.* à chaque itération, la configuration est modifiée de la façon suivante :
- ajout d'un nouvel objet,
- suppression d'un objet existant,
- translation d'un objet existant,
- rotation d'un objet existant,
- modification de la forme et du profil d'intensité d'un objet existant.

L'estimation est ensuite réalisée au sens du maximum *a posteriori*, *i.e.* la configuration de la chaîne de Markov qui maximise la densité *a posteriori* des paramètres du modèle est extraite.

Chapitre 3

Prétraitements des données

Sommaire

Ce chapitre introduit différentes techniques de prétraitement des données basées sur des approches par tests multiples. Dans une première partie, nous introduirons la nécessité de prétraiter les données dans le cas de la détection de sources à faible rapport signal à bruit. Nous illustrerons nos propos à l'aide d'un exemple sur le cube de données synthétiques DryRun. La seconde partie de ce chapitre est un état de l'art des méthodes de seuillages des données qui permettent de contrôler un critère d'erreur. Et enfin nous introduirons deux stratégies de prétraitement des données MUSE que nous avons implémentées en complément de la méthode de détection décrite dans le chapitre 2.

3.1 Prétraiter les données : une nécessité

Dans le chapitre 2, nous avons introduit une méthode de détection de sources dans les données hyperspectrales MUSE basée sur la représentation de la configuration de galaxies par un processus ponctuel marqué plongé dans un cadre bayésien. Ne disposant d'aucune connaissance *a priori* sur le nombre de sources, leurs positions, leurs formes géométriques ou encore leurs

intensités, nous avons introduit des *a priori* non informatifs dans le modèle bayésien conduisant à une absence de régularisation sur les positions et les intensités des objets. Si cette absence de pénalisation sur les intensités permet la détection de galaxies de très faible intensité, cela augmente le risque que le moindre pixel de bruit d'intensité significativement plus forte que la moyenne soit identifié comme une source par l'algorithme de détection afin de réduire les résidus de modélisation à cet endroit. Au lieu d'utiliser une régularisation dans la phase d'estimation bayésienne, nous préférons introduire une phase de prétraitement fréquentiste qui a pour but de guider la détection dans les zones du cube de données les plus probables. Nous parlerons de méthodes de seuillage puisqu'il s'agit de construire une liste de pixels du cube de données qui répondent à un critère de détection en opposition au reste des pixels qui seront classés comme pixels de bruit. Cette stratégie permettra également d'accélérer la convergence de l'algorithme d'échantillonnage RJMCMC introduit au chapitre 2 en restreignant les propositions d'objets lors des mouvements de naissance.

3.1.1 Bref rappel sur les tests multiples

Dans ce chapitre nous allons introduire des prétraitements basés sur des approches par tests multiples. Selon les approches, nous allons tantôt formuler des tests sur chacun des spectres formant le cube de données MUSE, tantôt tester chacun des pixels du cube. Dans le premier cas, le nombre de tests total N prendra la valeur $N \simeq 90000$, dans le second cas, $N \simeq 324 \times 10^6$. Dans les deux cas, le nombre de tests est conséquent et nécessite des approches adaptées aux données massives.

Nous avons résumé dans l'annexe A les notions de base de la problématique des tests multiples : définition du cadre théorique, contrôle du FWER, contrôle du FDR. Les différents prétraitements étudiés et présentés dans ce chapitre font appel à ces notions. Nous rappelons seulement dans le tableau 3.1 les notations du tableau A.2 introduit en annexe pour s'y référer rapidement dans la suite de ce chapitre.

Décision / Vérité	$\mathcal{H}_0$ est retenue	$\mathcal{H}_1$ est retenue	Total
$\mathcal{H}_0$ est vraie	N_0 - a	a	N_0
$\mathcal{H}_1$ est vraie	N_1 - b	b	N_1
Total	$N_0 + N_1$ - R	R	N

TABLEAU 3.1 – Rappel du tableau A.2 qui illustre la répartition des N tests en fonction de l'hypothèse $\mathcal{H}_i$ réelle et la décision prise à l'issu des tests. Le nombre de tests pour lesquels l'hypothèse $\mathcal{H}_0$ est vraie est noté N_0 et le nombre de tests pour lesquels l'hypothèse $\mathcal{H}_1$ est vraie est noté N_1. Le nombre $R = a + b$ correspond au nombre de cas où l'hypothèse $\mathcal{H}_0$ a été rejetée, avec a le nombre de cas où l'hypothèse $\mathcal{H}_0$ a été rejetée à tort et b le nombre de cas où l'hypothèse $\mathcal{H}_0$ a été rejetée à raison.

Dans le cas considéré dans ce paragraphe, c'est-à-dire la construction d'une carte de positions spatiales (p, q) qui contiennent probablement la contribution d'une source, le problème de test d'hypothèses peut se formuler ainsi :

— sous $\mathcal{H}_0$, l'échantillon ne contient que du bruit,
— sous $\mathcal{H}_1$ l'échantillon contient du signal et du bruit.

Finalement, décider pour chaque échantillon (spectre ou pixel) entre $\mathcal{H}_0$ et $\mathcal{H}_1$ revient à faire une classification à deux classes :

— la classe $\mathcal{C}_0$ des échantillons ne contenant que du bruit,
— la classe $\mathcal{C}_1$ des échantillons contenant du signal,

qui se traduit par un seuillage des données.

3.1.2 Réduire l'espace d'exploration des données

Il faut noter que d'un point de vue théorique, l'espace Ω des configurations d'objets qui tombent dans le sous espace borné $A \subset \mathcal{P}$ [1], où A représente le domaine spatial de MUSE (300×300pixels), est infini. Il faudrait donc explorer un ensemble infini de configurations afin de pouvoir estimer au sens du maximum *a posteriori* la configuration de galaxies. Pour restreindre l'exploration des configurations d'objets aux plus probables **nous avons fait le choix de prétraiter les données MUSE afin de dresser une liste de pixels qui appartiennent probablement à une source**, selon un critère à définir. **Cette liste de pixels sera utilisée afin de définir la fonction d'intensité $\lambda(\cdot)$ du processus ponctuel marqué** (voir définition à l'équation (B.4)).

Seuiller une image ou un cube de données avant d'appliquer les méthodes de détection de sources est une technique classique en astrophysique comme nous avons pu le voir dans le chapitre 1 avec les méthodes SExtractor, SFIND, DUCHAMP ou encore SoFiA. Cependant, dans le cadre de la détection de galaxies de très faible intensité, il y a un compromis à faire entre le choix d'un seuil suffisant pour limiter les fausses détections et abaisser suffisamment le seuil pour permettre la détection des sources d'intensités les plus faibles. Les approches par tests multiples que nous présenterons dans la suite de ce chapitre visent à proposer différents types de contrôle des erreurs lors de l'étape de seuillage des données.

3.1.3 Normalisation des données

Les cubes de données MUSE contiennent du bruit additif dont la moyenne et la variance varient en fonction de la longueur d'onde (nous avons fait l'hypothèse que le bruit est stationnaire spatialement, à condition que les données aient été correctement corrigées lors de la construction du cube). La première étape du prétraitement consiste à centrer et réduire les données afin de pouvoir formuler ensuite des tests sur des données normalisées. **Nous avons besoin de méthodes précises et robustes pour estimer la moyenne et la variance du bruit car les performances de détection de ces tests nécessitent que les données soient centrées et réduites.** Lors de l'étape de centrage et de réduction du bruit, les paramètres de moyenne et de variance doivent être estimés pour chaque image du cube. Dans ce cas, **les paramètres de moyenne et de variance sont supposés constants sur chaque feuillet de longueur d'onde** λ et doivent être estimés. Pour cela, il existe différentes approches :

— l'estimation paramétrique, où la densité de probabilité du processus aléatoire observé est supposée connue, et dépendant de paramètres que nous devons alors estimer,
— l'estimation non paramétrique où aucune hypothèse concernant la densité de probabilité n'est faite, les moments d'ordre 1 et 2 sont estimés uniquement à partir de l'ensemble des observations.

Dans le paragraphe 3.1.3.1, nous exposons le principe de l'estimation de la moyenne et de la variance par σ-clipping classique. Nous proposons dans le paragraphe 3.1.3.2 une approche par σ-clipping par point fixe fournissant une estimation plus précise et plus robuste à la présence de

1. Se référer à l'annexe B pour les notations sur les processus ponctuels marqués.

sources. Et enfin dans le paragraphe 3.1.3.3 nous rappelons la méthode d'estimation paramétrique au sens du maximum de vraisemblance sur données tronquées proposée par Efron [2010]. Les performances de ces trois méthodes seront mesurées sur des données synthétiques présentant différentes configurations de sources à différents rapports signal à bruit, les résultats sont donnés dans le paragraphe 3.1.3.4.

Il faut noter qu'avant tout traitement, le cube de données peut être divisé pixel par pixel par la racine carrée du cube de variance Σ_{MUSE} associé, voir la définition équation (1.11). Cette opération permet de ramener théoriquement la variance du bruit à l'unité, le cube de variance est estimé par les astrophysiciens lors de l'étape de moyennage des différentes poses individuelles. La division par ce cube de variance Σ_{MUSE}, lorsqu'il est jugé fiable par les astrophysiciens, est censée atténuer l'intensité des pixels présentant une forte variance entre les poses individuelles. Sur les données synthétiques que nous allons utiliser dans ce chapitre, cette étape ne sera pas effectuée, le bruit additif est généré selon une loi normale et il est i.i.d. spatialement et spectralement.

3.1.3.1 Estimation de la moyenne et de l'écart-type par la méthode de σ-clipping implémentée dans mpdaf.

Une méthode d'évaluation de la moyenne et de la variance des pixels d'une image par σ-clipping est implémentée dans la suite logicielle *mpdaf* associée aux données MUSE. Il s'agit ici d'une méthode non paramétrique puisqu'aucune hypothèse n'est faite concernant la distribution des pixels de bruit. La moyenne, ou la médiane qui est plus robuste aux valeurs anormales, et la variance sont estimées à l'aide d'un algorithme itératif qui tronque les données à chaque itération k à plus ou moins $3\hat{\sigma}_{k-1}$ autour de la moyenne (ou médiane) $\hat{m}_{k-1}$ et ajuste $\hat{m}_k$ et $\hat{\sigma}_k$ à l'aide de la nouvelle troncature. Nous rappelons le principe de l'algorithme d'estimation de la moyenne et de la variance, pour une longueur d'onde λ donnée, dans l'encadré 3.1. En partant de l'hypothèse que, sur chaque image, le nombre de pixels de bruit est largement supérieur au nombre de pixels contenant la contribution d'une source (ce qui est en pratique vérifié), alors $\hat{m}_k$ et $\hat{\sigma}_k$ sont proches de la moyenne et de la variance du bruit, et la normalisation ne porte bien que sur le bruit.

Encadré 3.1 – Méthode de σ-clipping *mpdaf*

Initialisation : I_0 = tous les pixels de l'image.
Pour k $\in \{1, \cdots, n_{iter}\}$ **:**

1. Calcul de la médiane m_k =médiane(I_{k-1})
2. Calcul de l'écart-type σ_k =std(I_{k-1})
3. Sélection des données telles que : $I_k = \{I_{k-1}(i) \text{ t.q. } I_{k-1}(i) \leqslant (m_k + 3\sigma_k)\}$

Résultat de l'estimation : $(\hat{m}, \hat{\sigma}) = (m_{n_{iter}}, \sigma_{n_{iter}})$

L'algorithme présente l'inconvénient de n'exclure des données utilisées pour réaliser l'estimation que les pixels dont l'intensité est très élevée (et positive) et conserve un grand nombre de pixels appartenant à des sources d'intensités moins fortes. De plus, le nombre d'itérations, en pratique $n_{iter} = 3$, est arbitraire (l'algorithme ne converge pas si n_{iter} augmente). La troncature à 3σ conserve, à chaque itération, 99.7% des données dans le cas de données gaussiennes pour réaliser l'estimation de la moyenne et de la variance. L'estimation de la moyenne et de la variance risque d'être biaisée par la présence de sources.

3.1.3.2 Estimation de la moyenne et de l'écart-type par la méthode de σ-clipping par point fixe

La méthode de σ-clipping par point fixe que nous proposons, contrairement à la méthode implémentée dans *mpdaf* (cf. paragraphe 3.1.3.1), fait l'hypothèse que le bruit est gaussien pour estimer l'écart-type de ce bruit. L'estimation repose toujours sur une troncature des données (*clipping*) pour estimer la moyenne et l'écart-type, mais ici, nous allons tenir compte de la distribution des données (loi gaussienne tronquée par le clipping pour éviter de prendre en compte les valeurs extrêmes dues aux sources). Par conséquent, il est possible de calculer les facteurs correctifs nécessaires à l'estimation de l'écart-type pour une loi gaussienne tronquée. Le principe de l'estimation par σ-clipping par point fixe est détaillé dans l'encadré 3.2. Nous introduisons les quantités suivantes :

- le premier et le troisième quartiles $Q1$ et $Q3$ tels que $\Phi(Q1) = 0.25$ et $\Phi(Q3) = 0.75$ où Φ est la fonction de répartition de la loi normale,
- le lien entre l'écart-type σ d'une loi gaussienne centrée et l'interquartile $IQR = Q3 - Q1$ est donné par la relation $\sigma(Q3 - Q1) = \sigma \times (\Phi^{-1}(0.75) - \Phi^{-1}(0.25)) = \sigma \times 1.349$,
- le facteur de troncature β, tel que $\kappa = \Phi^{-1}(\beta)$, permet de définir l'intervalle de valeurs conservées pour l'estimation : $[\mu - \kappa\sigma, \mu + \kappa\sigma]$, β est la proportion de valeurs conservées, de chaque côté de la médiane, typiquement $\beta = 0.9$ (et donc $\kappa = 1.28$),
- le facteur correctif $\mathcal{F}_1$ est introduit pour prendre en compte le fait que l'estimation de l'écart-type est réalisée à partir de l'interquartile des données gaussiennes tronquées à $\pm\Phi^{-1}(\beta)\sigma$.

Le calcul du facteur correctif $\mathcal{F}_1$ et les équations de l'algorithme du point fixe sont définis dans l'annexe F.

ENCADRÉ 3.2 – Méthode de σ-clipping par point fixe

Initialisation :

- I_0 = tous les pixels de l'image.
- Calcul de l'interquartile : IQR $= Q_3 - Q_1$
- Calcul de la médiane : m_0 =médiane(I_0)
- Calcul de l'écart-type : $\sigma_0 =$ IQR$/1.349$.
- Facteur de troncature à κ.
- Calcul du facteur correctif $\mathcal{F}_1$ pour l'estimation de l'écart-type.

Algorithme du point fixe, pour k = 1, $\cdots$, jusqu'à convergence :

1. Sélection des données telles que :

$$I_k = \{I_0(i) \text{ t.q. } \left|I_0(i) - m_{k-1}\right| \leqslant \kappa\sigma_{k-1}\}$$

2. Calcul de la médiane m_k
3. Calcul de l'écart-type

$$\begin{aligned} Q1_k &= \text{quartile}(I_k, 25\%) \\ Q3_k &= \text{quartile}(I_k, 75\%) \\ \sigma_k &= \frac{Q3_k - Q1_k}{\mathcal{F}_1} \end{aligned}$$

[Optionnel] Estimation fine de σ après convergence :
Soit k_f l'itération à laquelle l'algorithme du point fixe a convergé.

- $I_{clip} = \{I_0(i) \text{ t.q. } -3\sigma_{k_f} \leqslant I_0(i) - m_{k_f} \leqslant 0\}$
- Calcul du facteur $\mathcal{F}_2$ de correction pour la troncature entre $-3\sigma_{k_f} + m_{k_f}$ et m_{k_f}.
- Estimation de l'écart-type :

$$\sigma = \frac{\sqrt{\frac{1}{N_{I_{clip}}} \sum_{i=1}^{N_{I_{clip}}} \left(I_{clip} - m_{k_f}\right)^2}}{\mathcal{F}_2}$$

Résultat de l'estimation : $(\hat{m}, \hat{\sigma}) = (m_{n_{iter}}, \sigma)$

La méthode converge typiquement en quelques itérations ($k_f \simeq 10$). Dans cet algorithme du point fixe, les données sont tronquées à $\pm\kappa\sigma$ autour de la médiane, ce qui permet de s'affranchir des contributions des sources. Et sous l'hypothèse que la distribution du bruit sous $\mathcal{H}_0$ est bien symétrique, nous obtenons bien un estimateur de la moyenne, qui est aussi la médiane, du bruit. Le paramètre de troncature β pourrait être choisi beaucoup plus faible que dans le σ-clipping traditionnel car la troncature est symétrique, et l'estimation robuste de σ est corrigée ensuite par le facteur correctif $\mathcal{F}_1$ qui correspond au modèle paramétrique supposé (ici un modèle gaussien).

L'estimation optionnelle de l'écart-type, réalisée plus finement à la fin des k_f itérations, utilise les données comprises entre la médiane m_{k_f} et la médiane décalée de $-3\sigma_{k_f}$. L'idée étant que les données inférieures à la médiane ne sont pas censées être contaminées par les sources, et sous hypothèse de distribution gaussienne, donc de symétrie par rapport à la moyenne, il est possible de corriger la variance de la loi tronquée pour réaliser l'estimation fine de l'écart-type.

3.1.3.3 Estimation paramétrique de la moyenne et d'écart-type par maximum de vraisemblance

Supposons que l'hypothèse de bruit gaussien soit valide, estimer la moyenne et la variance du bruit peut alors se faire de façon paramétrique au sens du maximum de vraisemblance. L'estimation devient un problème d'optimisation classique. Afin d'éviter la contamination des données par les sources lors de l'estimation de la moyenne et de l'écart-type de la loi gaussienne utilisée pour modéliser les données sous l'hypothèse nulle, les données $I_0 = [I_0(1), \cdots, I_0(N)]$ sont tronquées, et la vraisemblance est exprimée pour ces données tronquées.

ENCADRÉ 3.3 – Estimation paramétrique au sens du maximum de vraisemblance

1. Initialisation :
   — $I_0 = [I_0(1), \cdots, I_0(N)]$ = tous les pixels de l'image.
   — Calcul de la médiane m_0 =médiane(I_0)
   — $B_{inf} = -4\sigma_0$
   — $B_{sup} = 2\sigma_0$
   — Calcul de l'écart-type $\sigma_0 = IQR/1.349$
2. Sélection des données telles que : $I_{clip} = \{I_0(i) \text{ t.q. } B_{inf} \leqslant I_0(i) - m_0 \leqslant B_{sup}\}$
3. Définition de la fonction de vraisemblance des données :

$$\mathcal{L}_{I_{clip}}(m, \sigma) = \frac{1}{\Phi(B_{sup}) - \Phi(B_{inf})} \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma}(I_{clip}(i) - m)^2\right)$$

4. Optimisation numérique de la log-vraisemblance :

$$(\hat{m}, \hat{\sigma}) = \text{argmax}\left(\log \mathcal{L}_{I_{clip}}\right)$$

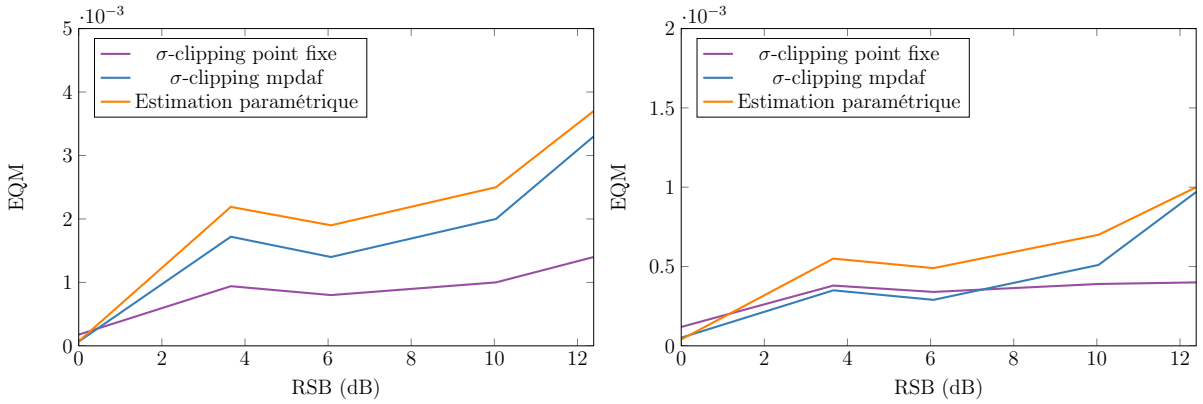
Résultat de l'optimisation : $(\hat{m}, \hat{\sigma})$.

Quelle que soit la méthode d'optimisation utilisée (Powell, gradient conjugué, etc) l'estimation des paramètres converge rapidement mais elle est sensible au choix des bornes, en particulier B_{sup}, pour la troncature des données. Le choix de B_{inf} n'est pas crucial, il permet seulement d'éviter les valeurs aberrantes très négatives, en revanche le choix de B_{sup} est important. Prendre B_{sup} trop faible (par exemple $B_{sup} = 0$) permet de rejeter une grande partie des données contaminées par les sources, mais rend l'estimation peu robuste, les données inférieures à la moyenne sont alors ajustées par des lois normales tronquées dont la moyenne et la variance sont surestimées. En pratique, la valeur $B_{sup} = 2\sigma$ est la borne qui fournit la meilleure estimation sur les données synthétiques.

3.1.3.4 Evaluation des performances des différentes méthodes

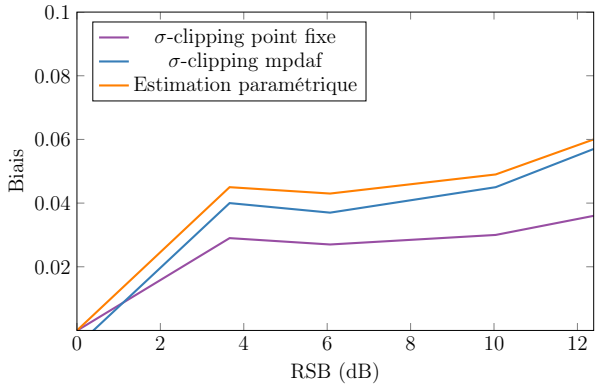
Afin d'évaluer les performances des différentes méthodes d'estimation de la moyenne et de la variance du bruit, nous générons six images de taille 150×150 contenant un certain nombre de sources (entre 5 et 40 selon les images). Pour des raisons de simplicité, les sources ont toutes le même profil spatial d'intensité (un profil gaussien en deux dimensions) et leurs intensités sont tirées de manière uniforme sur un intervalle de valeurs garantissant des RSB par objets similaires à ceux que l'on trouvent sur les données réelles. Nous obtenons 6 images différentes

dont le RSB moyen [2] varie entre 0 dB et 12.4 dB une fois le bruit gaussien i.i.d. de moyenne nulle et de variance unité ajouté. La figure 3.1 illustre les performances des différentes méthodes d'estimation de la moyenne et de la variance du bruit. L'estimation par σ-clipping par point fixe fournit des estimateurs de moyenne et de variance moins biaisé que les autres méthodes dès lors qu'il y a un certain nombre de sources présentes dans l'image ($RSB \geqslant 1$dB). L'estimation paramétrique est moins performante que le σ-clipping implémenté dans *mpdaf* pour un RSB inférieur à 12dB.

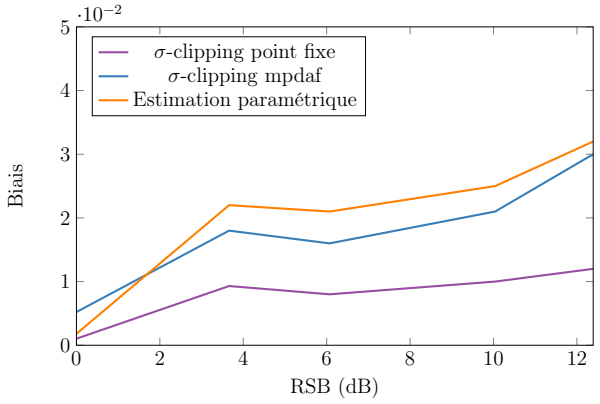


(a) EQM moyenne

(b) EQM écart-type



(c) Biais moyenne



(d) Biais écart-type

FIGURE 3.1 – Performances des différentes méthodes d'estimation de la moyenne et de l'écart-type sur des données synthétiques contenant entre 5 et 35 sources dans une image de taille 150 × 150 pixels. Les performances sont représentées en fonction du RSB moyen des images testées. Les biais et les EQM sont calculée empiriquement sur 3000 réalisations du bruit gaussien i.i.d.

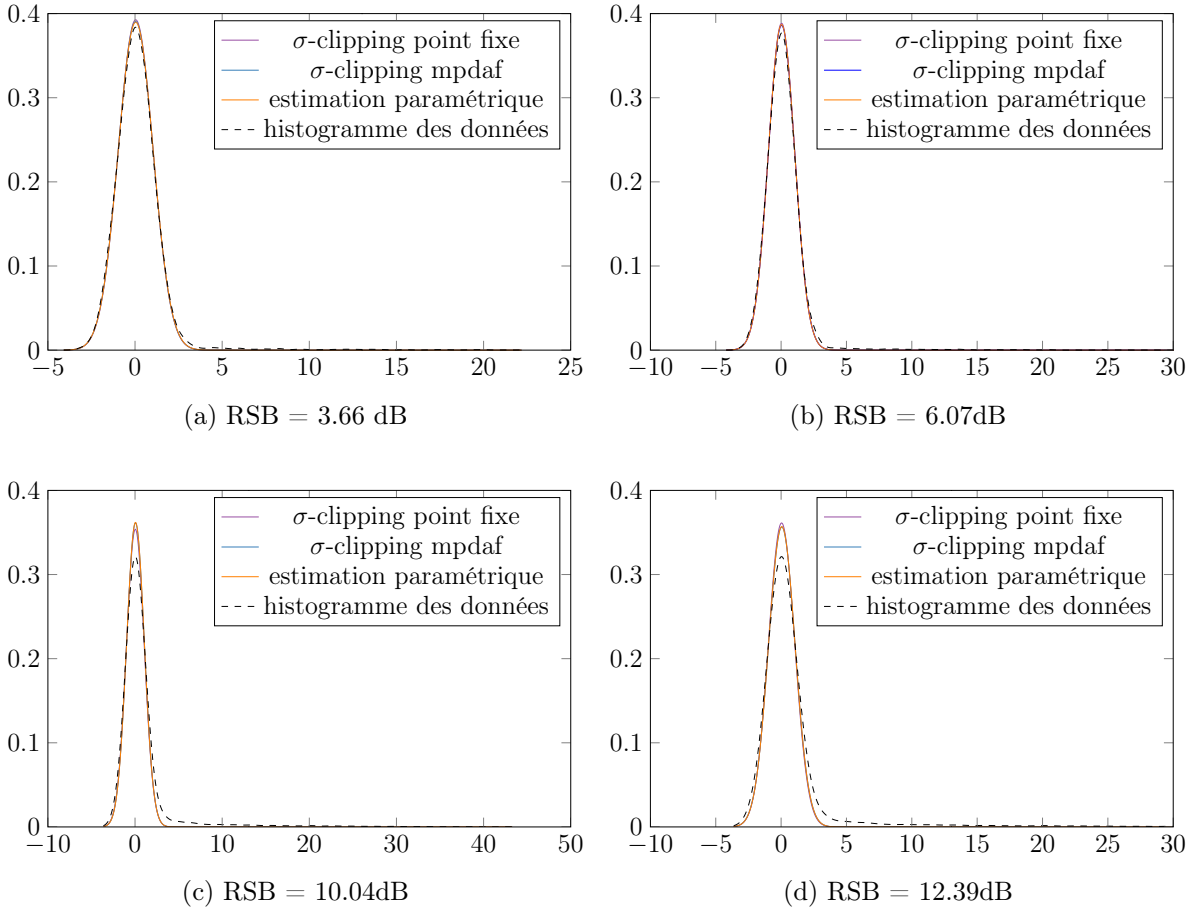
Le résultat des différentes méthodes d'estimation est présenté sur la figure 3.2 pour quatre images (de RSB 3.66dB, 6.07dB, 10.04dB et 12.39dB). Pour les images à faible RSB (3.66dB et 6.07dB) les trois méthodes d'estimation fournissent des résultats assez similaires, ce qui est confirmé par les courbes de performances présentées sur la figure 3.1. En revanche pour les images de plus fort RSB (10.04dB et 12.39dB), en zoomant sur le mode de l'histogramme, nous observons que l'estimation réalisée par σ-clipping de *mpdaf* et la méthode paramétrique (les courbes sont confondues) sont biaisées au niveau de l'estimation de la moyenne, la moyenne est légèrement surestimée. Nous pouvons également constater que l'estimation de la proportion π_0

2. Le RSB moyen d'une image $\mathcal{I}$ est défini par : $RSB = 20 \log\left(\frac{m_{\mathcal{I}}}{\sigma}\right)$ où $m_{\mathcal{I}}$ est la moyenne de l'image (avant ajout de bruit) et σ est l'écart-type du bruit additif gaussien.

de pixels vérifiant l'hypothèse nulle, qui permet d'estimer la composante $\pi_0 f_0$ du mélange de lois, où f_0 est la densité de probabilité sous $\mathcal{H}_0$, est biaisée également. Nous utilisons l'estimation conservative de Storey [2002] pour obtenir $\hat{\pi}_0$, ce qui conduit à l'expression :

$$\hat{\pi}_0 = \frac{\#\{X < \hat{m} | \mathcal{H}_0\}}{\frac{1}{2}N} = \min\left(\frac{2N_0}{N}, 1\right)$$

où N_0 est le nombre de pixels de valeur inférieure à la moyenne $\hat{m}$ estimée par σ-clipping par point fixe (estimateur le moins biaisé). La proportion $\hat{\pi}_0$ estimée est, par construction, surestimée, certains pixels d'intensité inférieure à la moyenne étant contaminés par des sources de faibles intensités (périphérie des galaxies, galaxies lointaines).



(a) RSB = 3.66 dB (b) RSB = 6.07dB

(c) RSB = 10.04dB (d) RSB = 12.39dB

FIGURE 3.2 – Histogramme des données (courbe pointillée noire) et distributions $\hat{\pi}_0 \hat{f}_0$ estimées à l'aide des différentes estimations de la moyenne et de la variance (courbes trait plein violette, bleue et orange).

3.1.3.5 Choix de la méthode de centrage et de réduction

Nous avons présenté dans les paragraphes précédents, trois méthodes d'estimation de la moyenne et de la variance des pixels sous l'hypothèse $\mathcal{H}_0$, deux méthodes classiques : le σ-clipping de *mpdaf* et l'estimation paramétrique et une méthode originale : le σ-clipping par point fixe. Étant données les performances sur données synthétiques des trois méthodes étudiées, et en considérant que les cubes de données MUSE contiendront toujours un certain nombre de sources brillantes, à spectres continus, de l'ordre de quelques dizaines, nous avons choisi de conserver la

méthode de σ-clipping par point fixe qui fournit l'estimation la moins biaisée pour la moyenne et la variance ainsi que la plus faible erreur quadratique moyenne. Cette méthode fournit en effet un estimateur robuste et précis des paramètres de moyenne et variance du bruit en réalisant l'estimation sur les données tronquées, ce qui élimine en grande partie les pixels contaminés par des sources.

3.1.4 Filtrage adapté, pour quoi faire et avec quelles conséquences ?

Le prétraitement que nous souhaitons mettre en place vise en particulier à améliorer la détectabilité des galaxies de plus faibles intensités et à favoriser la recherche de galaxies dans les zones les plus probables. En présence de bruit, ces galaxies sont particulièrement difficiles à détecter, notamment lorsque la raie Lyα est fortement décalée dans le rouge, au niveau des résidus de soustraction des raies du ciel. Lorsque la réponse de la source que l'on cherche à détecter est connue, le filtrage optimal, au sens de la maximisation du RSB de la source, est le filtrage adapté. La réponse impulsionnelle du filtre adapté à une source est la réponse de cette source retournée (ici dans l'espace à trois dimensions).

Dans le cas de MUSE, les galaxies de faible intensité sont également des galaxies lointaines, qui sont non résolues ou faiblement résolues spatialement, et dont le spectre est composé principalement d'une raie d'émission, la raie Lyα. La réponse de ce type de galaxie doit donc être assez proche de la PSF en trois dimensions de l'instrument. Les galaxies à raie Lyα font encore l'objet d'études et de modélisations de la raie Lyα (voir par exemple Garel [2011]). Les astrophysiciens disposent de modèles astrophysique de la raie Lyα. Suleiman et al. [2013] ont proposé de modéliser les 10^4 modèles de raie Lyα fournis par les astrophysiciens par un profil spectral estimé par approche minimax. Dans les travaux de Paris et al. [2013b], ce profil est étudié ainsi que deux autres approches : le profil estimé par SVD et 7 profils moyens distincts obtenus par K-SVD censé représenter au mieux les 10^4 modèles de raie Lyα. Il ressort de ces travaux que les modèles de raies Lyα ont tous une petite extension spectrale (sur une dizaine de bandes) avant composition par la LSF. Le spectre d'une galaxie de type émetteur Lyα, observé avec MUSE, sera donc le résultat de la convolution de la LSF de l'instrument avec une composante spectrale non nulle sur quelques bandes spectrales consécutives (un nombre inférieur à 10 bandes avec la résolution de MUSE).

Sur la base des résultats obtenus par Suleiman et al. [2013] et Paris et al. [2013b], seul l'élargissement de la raie Lyα sera prise en compte en choisissant un filtrage adapté à la LSF élargie. Nous allons adopter l'approximation suivante :

> **Approximation :** *La réponse en trois dimensions d'une galaxie lointaine est modélisée par la PSF de l'instrument spectralement élargie centrée sur les coordonnées (p, q, λ), où (p, q, λ) sont les coordonnées (en pixels) du centre de la galaxie considérée.*

La réponse en trois dimensions de la PSF de l'instrument MUSE est modélisée en chaque point (p, q, λ) du cube, il est donc possible de réaliser le filtrage du cube de données adapté à la PSF spectralement élargie. Pour élargir la LSF de MUSE, il suffit de convoluer la LSF de MUSE avec un profil symétrique, non nul sur quelques bandes spectrales consécutives, comme illustré sur la figure 3.3.
Ce filtrage adapté est sous-optimal[3], mais ne connaissant pas *a priori* la forme des raies Lyα

3. Notons également que le filtrage appliqué ici n'est pas un filtrage adapté au sens strict du terme, d'une part le filtre n'est pas invariant par translation dans la dimension spectrale, d'autre par, le bruit, supposé i.i.d. dans le modèle, ne l'est pas dans la pratique. Ceci conduit à un filtrage sous-optimal. De plus, sur certains jeux de données réelles, un étape de réduction du cube de données par le cube de variance estimée peut être appliquée avant le filtrage adapté, ce qui peut déformer le profil des raies, notamment si elles existent dans les grandes longueurs d'onde où subsistent des résidus de la soustraction des raies du ciel. Notons cependant qu'empiriquement le filtrage adapté tel qu'il est décrit dans ce manuscrit, améliore considérablement le RSB des sources quasi-ponctuelles.

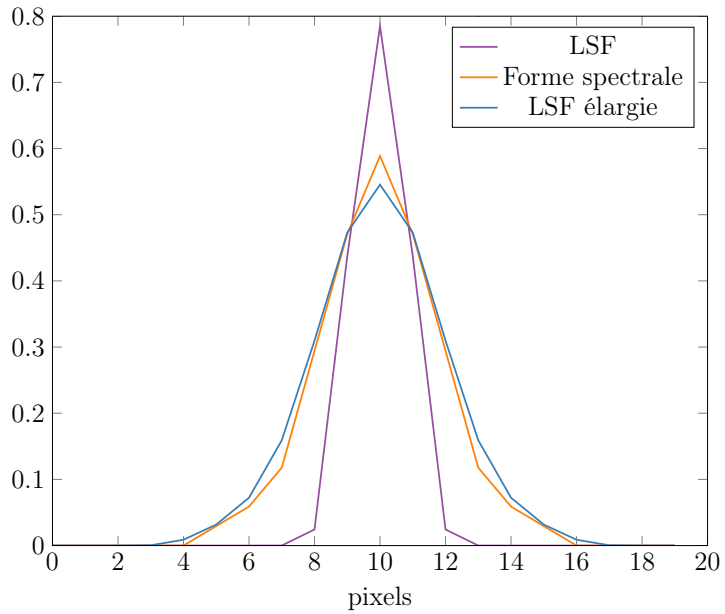


FIGURE 3.3 – Elargissement de la LSF de MUSE par un profil spectral symétrique, contenant peu d'informations sur la forme de la raie d'émission recherchée. Tous les profils sont normalisés avec une norme ℓ_2.

observées dans un cube de données, et étant donnée la dimension des données, il n'est pas raisonnable, en terme de coût, d'envisager réaliser un très grand nombre de filtrages adaptés. Il pourrait être intéressant d'utiliser un des profils moyens estimés par Suleiman et al. [2013] et Paris et al. [2013b] afin d'élargir la LSF pour remplacer le profil symétrique que nous utilisons actuellement.

3.1.4.1 Définition mathématique

La FSF, notée F et la LSF spectralement élargie, notée $\widetilde{L}$, étant symétriques, la réponse impulsionnelle du filtre adapté à la PSF de MUSE est la PSF, spectralement élargie, elle-même. En reprenant les hypothèses établies dans le paragraphe 1.5.1 concernant la modélisation de la PSF de MUSE, le filtrage adapté en trois dimensions s'écrit :

$$\mathbf{Y}^f(p,q,\lambda) = (\mathbf{Y} \otimes h)(p,q,\lambda) = \sum_{\mu} \widetilde{L}_\mu(\lambda) \Big\{ \Big(F_\mu * \mathbf{Y}(\cdot,\cdot,\mu) \Big)(p,q) \Big\}. \tag{3.1}$$

Il faut noter que contrairement au chapitre précédent, la FSF est variable en longueur d'onde. Nous aurons besoin dans la suite de ce chapitre d'une notation vectorielle du filtrage adapté décrit par l'équation 3.1. Cette équation peut se reformuler simplement sous forme de vecteurs :

$$\mathbf{Y}^{f,(v)} = \mathbf{H}^T \mathbf{Y}^{(v)}, \tag{3.2}$$

où le vecteur $\mathbf{Y}^{f,(v)}$ de taille $N \times 1$, avec $N = P \times Q \times \Lambda$, est la sortie du filtrage adapté sous forme vectorisée, la matrice $\mathbf{H}$ de taille $N \times N$ est la matrice contenant dans chaque colonne $h_{p,q,\lambda}$ la réponse de la PSF centrée en un point (p,q,λ) du cube, sous forme vectorisée. Le vecteur $\mathbf{Y}^{(v)}$ de taille $N \times 1$ est la forme vectorisée du cube de données $\mathbf{Y}$ tel que $\mathbf{Y}^{(v)} = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_\lambda, \cdots, \boldsymbol{y}_\Lambda]^T$ avec $\boldsymbol{y}_\lambda$ un vecteur ligne de taille $(1 \times (P \times Q))$ représentant la vectorisation de l'image à la longueur d'onde λ du cube de données. La construction de la matrice $\mathbf{H}$ à partir de la modélisation de la PSF en deux composantes spatiale (FSF) et spectrale (LSF) est détaillée dans l'annexe E.

3.1.4.2 Exemple

La nécessité de mettre en place une phase de prétraitement basée sur le filtrage adapté à la PSF de l'instrument pour améliorer la détectabilité des galaxies lointaines de faible intensité sera illustrée dans ce paragraphe avec le cube de données synthétiques DryRun. Les galaxies lointaines de type Lyα sont caractérisées par la présence d'une unique raie dans leur spectre, nous pouvons donc étudier la valeur maximale de chaque spectre du cube pour caractériser la détectabilité des objets. Les autres galaxies du DryRun présentent soit une ou plusieurs raies d'émission de forte intensité, soit une composante continue significativement supérieure au niveau de bruit (voir tableau 1.1 pour le détail de la composition du cube), la valeur maximale de leur spectre sera donc également un bon indicateur de leur détectabilité. La projection du cube selon la valeur maximale de chaque spectre est représentée sur la figure 3.4 pour différents cas :

(a) Les données[4] sont seulement centrées-réduites à l'aide d'une estimation de la moyenne et de la variance par σ-clipping par point fixe à chaque longueur d'onde.

(b) Les données centrées-réduites (a) sont filtrées à l'aide du filtre dont la réponse impulsionnelle est la FSF de MUSE associée à ce cube. A noter que la FSF, qui varie spectralement, est considérée constante dans le champ d'observation, il s'agit donc d'une convolution spatiale plan par plan avec la FSF définie pour chaque longueur d'onde considérée. La FSF est normalisée : $\|F_\lambda\|_2^2 = 1$ pour tout $\lambda = 1, \cdots, \Lambda$.

(c) Les données centrées-réduites (a) sont filtrées à l'aide du filtre dont la réponse impulsionnelle est la LSF, spectralement élargie, de MUSE. Il s'agit ici d'une composition des spectres par l'opérateur de Fredholm puisque $L_\mu(\lambda) \neq L(\lambda - \mu)$. La LSF est normalisée : $\|L_\lambda\|_2^2 = 1$ pour tout $\lambda = 1, \cdots, \Lambda$.

(d) Les données centrées-réduites (a) sont filtrées à l'aide du filtre dont la réponse impulsionnelle est la réponse retournée de la PSF de MUSE associée à ce cube de données.

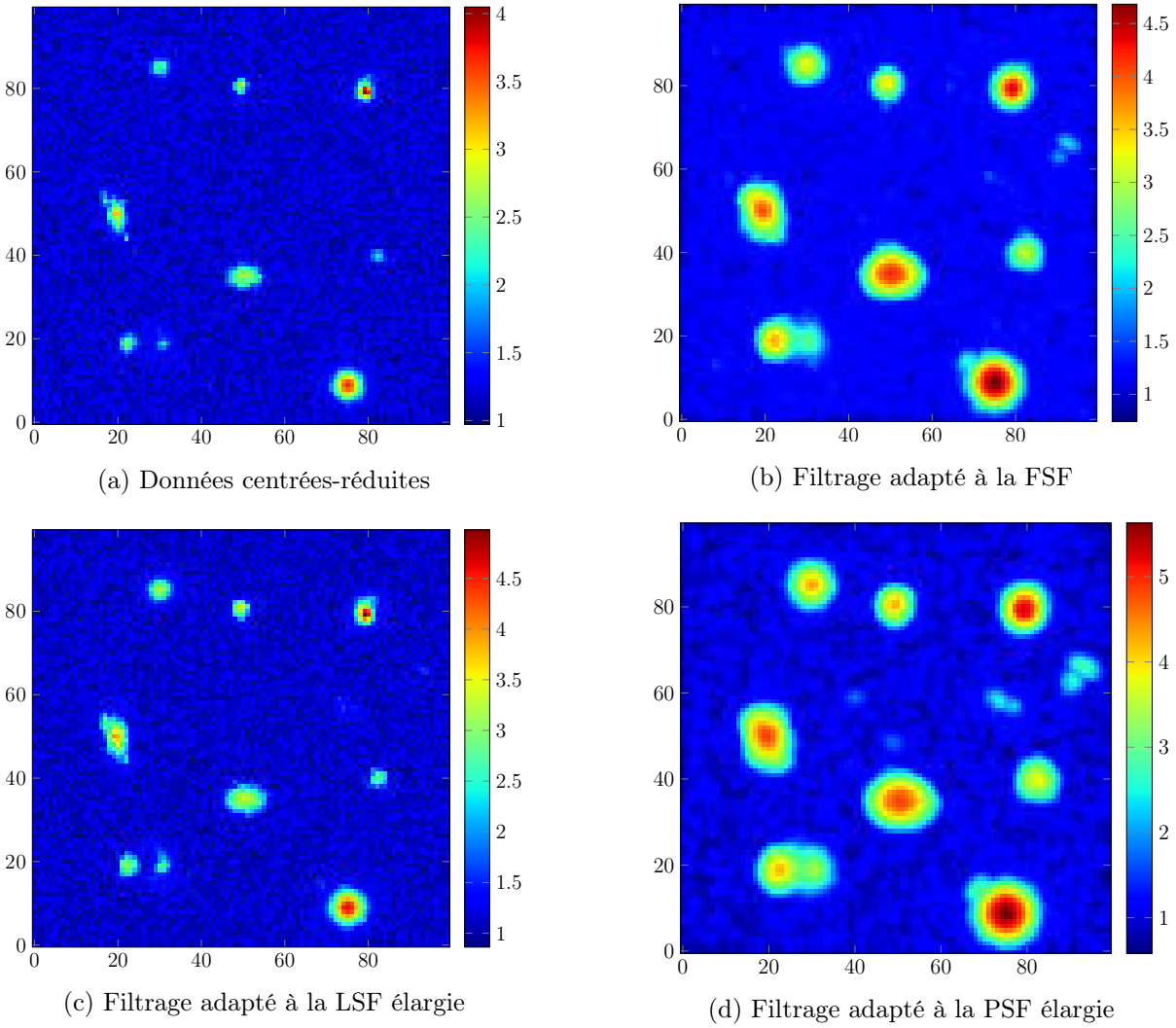
L'image $\mathcal{I}$ utilisée pour visualiser ces traitements est définie par :

$$\mathcal{I}(p,q) = \max_{\lambda=1\cdots\Lambda} \tilde{\mathbf{Y}}(p,q,\lambda), \tag{3.3}$$

où $\tilde{\mathbf{Y}}$ représente le cube de données après filtrage (a), (b), (c) ou (d).

La connaissance de ce cube de données nous permet d'évaluer pour différents seuils le nombre d'objets détectables grâce aux prétraitements. Avec la figure 3.5, nous illustrons l'utilité d'un filtrage adapté à la PSF de l'instrument pour améliorer la détectabilité des objets. Les courbes correspondent à la quantité $\frac{b}{N_1}$ (*c.f.* tableau 3.1) pour différentes valeurs de seuil et pour les quatre cas présentés sur la figure 3.4. Plus la quantité $\frac{b}{N_1}$ est élevée pour une même valeur de seuil, plus le filtrage utilisé permet l'amélioration de la détectabilité des objets. L'analyse de ces courbes doit être réalisée en regard des erreurs de décision menant à classer des pixels de bruit dans la classe des sources (voir figure 3.6 où est représentée la quantité $\frac{a}{R}$ (*c.f.* tableau 3.1) qui caractérise la proportion de ces pixels mal détectés parmi tous les pixels supérieurs au seuil) et vis-à-vis du nombre d'objets que les différents filtrages permettent de détecter pour une valeur de seuil donnée (voir figure 3.7). Sur la figure 3.6, nous sommes en train d'introduire implicitement la notion de proportion de fausses découvertes (FDP) qui est décrite plus en détail en annexe, dans le paragraphe A.2.2.

4. Dans le cas des données réelles, lorsque le cube de variance fourni avec les données est jugé fiable par les astrophysiciens, les données peuvent être au préalable divisé par la racine carré de ce cube de variance. Cette opération n'est pas réalisée pour le cube de données synthétiques DryRun.



(a) Données centrées-réduites

(b) Filtrage adapté à la FSF

(c) Filtrage adapté à la LSF élargie

(d) Filtrage adapté à la PSF élargie

FIGURE 3.4 – Affichage de la valeur maximale $\mathcal{I}(p,q)$ de chaque spectre du cube à la position (p,q) sur les données (a) non filtrées, (b) après filtrage adapté à la FSF seule, (c) après filtrage adapté à la LSF seule, (d) après filtrage adapté à la PSF totale (FSF + LSF). Les données ont été centrées-réduites au préalable, et l'échelle de couleur correspond à l'intensité de la valeur maximale de chaque spectre après traitement. La réponse impulsionnelle de chaque filtre est normalisée (norme ℓ_2).

L'analyse des courbes de performances présentées sur les figures 3.5, 3.6 et 3.7 nous permet de mettre en évidence les points suivants :

1. La détection des objets de plus faibles intensités, basée sur un critère de valeur maximale du spectre après différents types de filtrages, sera plus puissante avec un filtrage adapté à la PSF globale (FSF + LSF élargie) que sans filtrage ou un filtrage adapté dans l'une des deux dimensions (spatiale ou spectrale). Notons que le résultat n'est pas optimal puisque le filtrage adapté a été construit à partir d'une approximation de la FSF et de la LSF de l'instrument MUSE. De plus le bruit n'étant pas i.i.d. le résultat est sous-optimal, le filtrage permet cependant d'améliorer fortement le RSB des sources.
2. Sans filtrage adapté, seules les sources très brillantes, avec, éventuellement, une composante spectrale continue, sont détectables pour une proportion de fausses découvertes

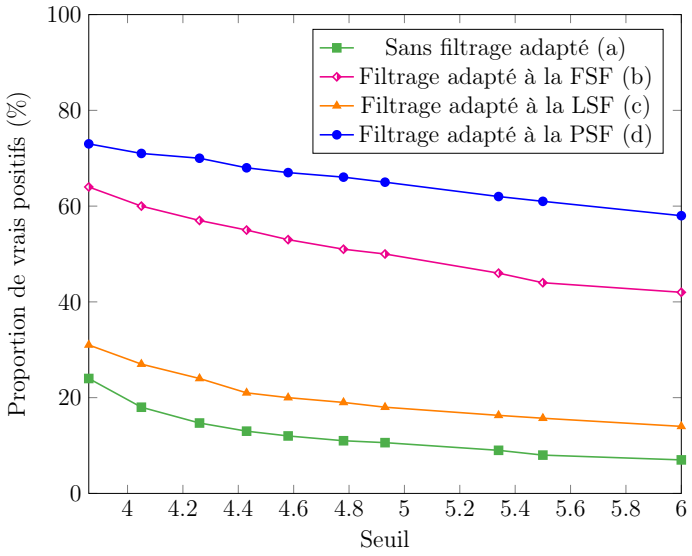


FIGURE 3.5 – Comparaison de la proportion $\frac{b}{N_1}$ de pixels appartenant à une source, supérieurs au seuil de décision parmi tous les pixels appartenant à une source pour différents types de filtrages des données : sans filtrage (courbe verte), avec filtrage adapté à la FSF (courbe magenta), avec filtrage adapté à la LSF (courbe orange) et avec filtrage adapté à la PSF globale (courbe bleue).

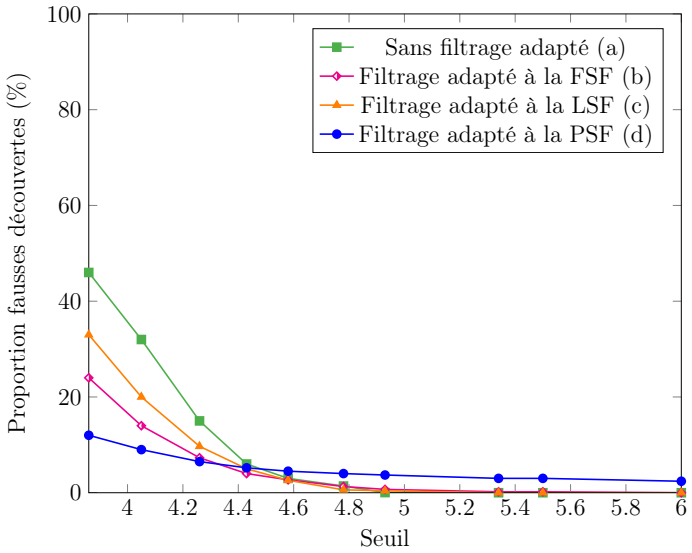


FIGURE 3.6 – Comparaison de la proportion $\frac{a}{R}$ de pixels ne contenant que du bruit et qui sont supérieurs au seuil de décision parmi tous les pixels supérieurs au seuil pour différents types de filtrages des données : sans filtrage (courbe verte), avec filtrage adapté à la FSF (courbe magenta), avec filtrage adapté à la LSF (courbe orange) et avec filtrage adapté à la PSF globale (courbe bleue).

raisonnable. Sur la figure 3.6, la courbe verte correspondant au cas (a) sans filtrage montre que pour le plus petit seuil choisi, 50% des spectres qui ont une valeur maximale supérieure au seuil sont en réalité des fausses détections (choix de $\mathcal{H}_1$ alors que $\mathcal{H}_0$ est vraie), et quand bien même nous accepterions que la moitié des détections soient erronées, le nombre d'objets détectables représenté par la courbe verte de la figure 3.7 pour cette

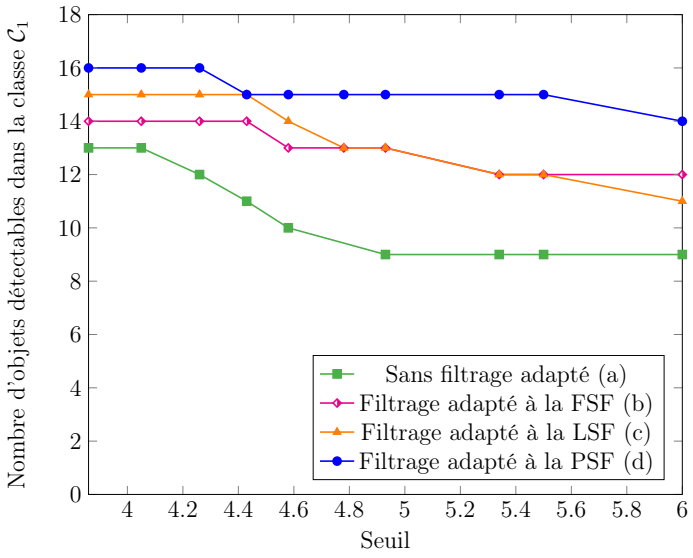


FIGURE 3.7 – Comparaison du nombre d'objets détectables à différentes valeurs de seuil pour différents types de filtrages des données : sans filtrage (courbe verte), avec filtrage adapté à la FSF (courbe magenta), avec filtrage adapté à la LSF (courbe orange) et avec filtrage adapté à la PSF globale (courbe bleue).

valeur de seuil est de 13 objets sur 18. En menant la même analyse dans le cas du filtrage adapté à la PSF (courbes bleues sur les figures 3.6 et 3.7), pour la plus petite valeur de seuil, seuls 12% des détections sont erronées et 16 sur les 18 objets sont détectables.

3. Si les performances du filtrage adapté à la PSF illustrées par la figure 3.6 (courbe bleue) semblent montrer que la proportion de fausses découvertes est plus élevée dans ce cas que pour les autres filtrages pour des grandes valeurs de seuil, il faut prendre en compte le fait que les autres seuillages sont moins puissants, c'est-à-dire que le nombre de pixels supérieurs au seuil est bien plus faible. Dans ces cas, la quasi totalité de ces pixels, classés dans $\mathcal{C}_1$ par le seuillage, appartiennent à une source et ont été détectés en tant que tels, mais, en réalité, seule une petite proportion des pixels appartenant à des sources ont été détectés avec ce seuil (voir figure 3.5).

Sur ce cube de données synthétiques, le filtrage adapté à la PSF spectralement élargie permet d'améliorer de façon significative la détectabilité des galaxies les moins brillantes. Nous utiliserons donc par la suite ce filtrage pour construire la carte de proposition des objets afin de s'assurer que cette carte de proposition contienne le plus grand nombre possible d'objets.

3.1.5 Du seuillage des données à la proposition des objets

Avec l'exemple donné dans le paragraphe 3.1.4.2, nous avons vu que le filtrage adapté permet d'améliorer la détectabilité des objets. Il est possible de définir un test T sur des échantillons du cube (pixels ou spectres) visant à décider si l'échantillon contient la contribution d'une source ($\mathcal{H}_1$) ou s'il ne contient que du bruit ($\mathcal{H}_0$). Si la loi de la statistique du test T peut être définie de manière analytique ou approchée de manière empirique, il est alors possible de relier la valeur du test à une quantité statistique qui peut être une p-valeur ou une probabilité de fausse alarme. Le résultat des tests appliqués aux échantillons est une carte binaire, qui répartit en deux classes $\mathcal{C}_0$ et $\mathcal{C}_1$ les positions spatiales (p, q) du cube.

Rappelons que l'un des buts des prétraitements présentés dans ce chapitre est de guider la proposition des centres des objets lors des mouvements de naissance de l'algorithme d'échan-

tillonnage RJMCMC introduit au chapitre 2. La carte binaire produite à l'issue du seuillage des données permet une première restriction de l'espace à explorer. Il est en effet inutile de proposer un centre d'objet sur un pixel qui a été classé dans $\mathcal{C}_0$. De plus, nous disposons d'un *a priori* fort sur le profil d'intensité spatial des galaxies : l'intensité est plus forte au centre et décroit en s'éloignant du centre. Tous les pixels de la classe $\mathcal{C}_1$ ne présentent pas le même intérêt lors de la proposition des centres. Nous verrons comment construire une carte de proposition tenant compte de l'*a priori* sur le pic d'intensité au centre de la galaxie selon les procédures de tests choisies.

3.2 État de l'art des approches par tests multiples pour les données MUSE

Les principales définitions et propriétés liées aux tests multiples sont résumées dans l'annexe A. Nous allons nous intéresser dans cette partie aux méthodes de la littérature qui peuvent se séparer en deux catégories de contrôles distincts : le contrôle des événements rares dans les spectres avec un contrôle de type FWER, et le contrôle de type FDR appliqué aux pixels du cube.

3.2.1 Contrôle des événements rares sur les spectres

Dans les paragraphes suivants, nous allons nous intéresser à un contrôle de type FWER (voir annexe A.2.1) pour la détection de raies d'émission dans les spectres du cube de données MUSE. Ce problème peut se formuler de la façon suivante : un spectre est composé de Λ échantillons qui sont, sous $\mathcal{H}_0$, distribués selon la même loi, et supposés indépendants. En présence d'une raie d'émission (resp. d'une composante spectrale continue), une faible (resp. une grande) proportion des échantillons du spectre sont en moyenne significativement plus élevés que les échantillons sous $\mathcal{H}_0$. La décision doit donc se faire entre les deux hypothèses suivantes :

— tous les échantillons sont distribués selon $\mathcal{H}_0$,
— au moins un échantillon (ou une proportion des échantillons) est distribué selon $\mathcal{H}_1$.

Nous nous intéresserons dans les paragraphes suivants au cas particulier de la détection des spectres à raie d'émission, sans composante continue. Dans la littérature, le critère de Higher Criticism (HC) a été proposé Tukey [1976] et développé par Donoho and Jin [2004] pour exercer un contrôle qui ressemble au critère de FWER. Au lieu de contrôler la probabilité de rejeter au moins une hypothèse nulle parmi les N tests, ce critère s'intéresse à la probabilité de rejeter une petite fraction d'hypothèses nulles. Les auteurs ont montré que le critère du HC est le test asymptotiquement le plus puissant pour la détection d'événements rares, ce qui présente un intérêt particulier pour notre problème de détection des raies Lyα dans les spectres des cubes de données MUSE. Le critère du HC ainsi que les améliorations proposées dans la littérature sont détaillés dans les paragraphes suivants, nous appliquerons les différents critères aux données synthétiques DryRun.

3.2.1.1 Le critère du Higher Criticism

Le critère du Higher Criticism (HC) a été proposé pour la première fois par Tukey [1976] dans ses notes de cours de statistiques à Princeton comme une procédure de contrôle global des erreurs dans le cas de tests multiples. Au lieu de contrôler indépendamment un nombre N de tests à un niveau $\alpha = 5\%$, il propose de s'intéresser au critère global $HC_{0.05,N}$ défini par :

$$HC_{0.05,N} = \frac{\sqrt{N}\,[\text{Fraction significative à } 0.05 - 0.05]}{\sqrt{0.05 \times 0.95}},$$

qui permet d'accepter ou de rejeter l'union des hypothèses nulles pour les N tests au lieu de considérer chaque test séparément. La contribution de Tukey s'est arrêté à cette proposition pour un taux $\alpha = 5\%$.

Le contrôle du critère HC a ensuite été repris et développé par Donoho and Jin [2004] pour la détection de signaux rares et de faibles intensités noyés dans du bruit pour un niveau α variant de 0 à un niveau maximum α_0 choisi par l'utilisateur :

$$HC_N^* = \max_{0<\alpha<\alpha_0} \frac{\sqrt{N}\,[\text{Fraction signifiative à } \alpha - \alpha]}{\sqrt{\alpha \times (1-\alpha)}}, \tag{3.4}$$

Par signaux *rares*, nous entendons que la contribution de ces signaux se restreint à quelques échantillons de l'observation (vecteur, matrices ou tableaux à plus grande dimension). Il s'agit donc de tester avec ce critère HC^* s'il existe une petite fraction d'hypothèses nulles $\mathcal{H}_0^i$ rejetées parmi les N tests effectués. Le critère du HC^* a été introduit pour un ensemble de variables $x_i \sim \mathcal{N}(\mu_i, 1)$ indépendantes dont la moyenne $\mu_i = 0$ sauf pour une petite fraction $\boldsymbol{\epsilon}$ pour lesquelles $\mu_i > 0$. Chaque variable x_i est distribuée selon l'une des deux hypothèses suivante

$$\left\{ \begin{array}{lcl} \mathcal{H}_0^i & : & \mu_i = 0 \\ \mathcal{H}_1^i & : & \mu_i > 0 \end{array} \right. \tag{3.5}$$

et le critère du HC^* doit permettre de décider si l'ensemble $\boldsymbol{x} = [x_1, \cdots, x_N]$ est distribué selon l'une ou l'autre des hypothèses suivantes :

$$\left\{ \begin{array}{lcl} \mathcal{H}_0 & : & x_i \sim \mathcal{H}_0^i \\ \mathcal{H}_1 & : & x_i \sim (1-\epsilon)\mathcal{H}_0^i + \epsilon \mathcal{H}_1^i \end{array} \right. \quad \text{pour tout } 1 \leqslant i \leqslant N$$

Le critère peut être reformulé en terme de p-valeurs :

$$HC_N^* = \max_{0<i<\alpha_0 N} \frac{\sqrt{N}\left[\frac{i}{N} - p_{(i)}\right]}{\sqrt{p_{(i)}(1-p_{(i)})}}, \tag{3.6}$$

avec $p_{(i)}$ les p-valeurs p_i associées aux variables x_i triées dans l'ordre croissant.

Pour conduire un test de niveau α sur l'ensemble des variables x_i, il faut trouver une valeur de seuil $h(N, \alpha)$ telle que :

$$Pr(HC_N^* > h(N,\alpha) | \mathcal{H}_0) \leqslant \alpha.$$

L'analyse asymptotique menée par Donoho and Jin [2004] montre que la procédure de seuillage à l'aide du HC^* est capable de séparer parfaitement les deux hypothèses (pour $N \to +\infty$) si l'on se situe au dessus de la frontière de détection $r > \rho^*(\beta)$ avec :

$$\rho^*(\beta) = \left\{ \begin{array}{ll} \beta - \frac{1}{2} & \frac{1}{2} < \beta \leqslant \frac{3}{4} \\ (1 - \sqrt{1-\beta})^2 & \frac{3}{4} < \beta < 1 \end{array} \right. ,$$

avec

$$\begin{array}{rcl} \mu_i & = & \sqrt{2r\log(N)} \text{ sous } \mathcal{H}_1^i \\ \boldsymbol{\epsilon} & = & N^{-\beta} \end{array} .$$

Si les données sont distribuées dans la région de détectabilité $r > \rho^*(\beta)$, alors il est possible de trouver une fonction $h(N, \alpha)$ qui dépend du nombre de tests et de la valeur de contrôle α, par exemple en approchant la loi du critère HC_N^* par méthode de Monte Carlo.

Dans les cas de détection les plus difficiles ($r < \frac{1}{4}$) les auteurs proposent un raffinement du critère HC_N^* en considérant que l'information permettant d'identifier les cas où $\mathcal{H}_1^i$ est vraie n'est pas localisée dans les valeurs extrêmes, *i.e.* la plus petite p-valeurs n'est pas forcément

significative pour la détection de $\mathcal{H}_1^i$. Le critère 3.6 est légèrement modifié afin de tenir compte de ce constat :

$$HC_N^+ = \max_{\substack{1<i\leqslant\frac{N}{2}\\ p_{(i)}\geqslant\frac{1}{N}}} \frac{\sqrt{N}\left[\frac{i}{N}-p_{(i)}\right]}{\sqrt{p_{(i)}(1-p_{(i)})}}, \tag{3.7}$$

3.2.1.2 Innovated Higher Criticism

Le critère du HC^* a été défini pour des tests indépendants, Hall et al. [2010] proposent une nouvelle version de ce critère, le iHC, pour seuiller un ensemble de données $\boldsymbol{x} = [x_1, \cdots, x_N] \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ avec Σ la matrice de covariance du bruit qui est supposée connue. Les hypothèses sur la moyenne μ_i de chaque variables x_i sont les mêmes que celles qui ont été définies par l'équation 3.5.

Hall et al. [2010] avancent les arguments suivants :

— le critère du HC^* ne tient pas compte des corrélations, *i.e.* si on permute les valeurs x_i du vecteur $\boldsymbol{x}$, le résultat restera inchangé,

— le critère du iHC^* est sensible aux corrélations (et donc aux permutations), il devrait être plus puissant que le critère HC^* dans le cas de données corrélées.

Ce critère iHC^* est asymptotiquement plus puissant pour la détection de signaux rares dans un bruit corrélé. La procédure proposée par Hall et al. [2010] est résumée dans l'encadré 3.4.

ENCADRÉ 3.4 – Calcul du iHC^*

1. Décomposition de Cholesky de la matrice de covariance : $\Sigma = LL^T$
2. Calcul de l'inverse de L : $U_N = L^{-1}$ telle que $U_N \Sigma U_N^T = \mathbf{I}_N$.
3. Construction d'une matrice *bande* $\tilde{U}_N(b_N)$ de bande passante b_N telle que :

$$(\tilde{U}_N)_{k,j} = \begin{cases} (U_N)_{k,j} & \text{si } k - b_n + 1 \leqslant j \leqslant k \\ 0 & \textit{sinon} \end{cases}$$

4. Normalisation de chaque colonne de $\tilde{U}_N(b_N)$ par leur norme ℓ_2 pour former la matrice $\bar{U}_N$.
5. Construction de la matrice $V_N = \bar{U}_N U_N$.
6. Application du critère iHC^* au vecteur $V_N\boldsymbol{x}$:

$$iHC_N^* = \max_{\frac{1}{N}\leqslant p_{(i)}\leqslant\frac{1}{2}} \frac{\sqrt{N}\left[\frac{i}{N}-p_{(i)}\right]}{\sqrt{p_{(i)}(1-p_{(i)})}}, \tag{3.8}$$

où les $p_{(j)}$, $j = 1, \cdots, N$, sont les p-valeurs triées dans l'ordre croissant des échantillons du vecteur $V_N\boldsymbol{x}$.

3.2.1.3 Application du HC^* et du HC^+ aux données DryRun

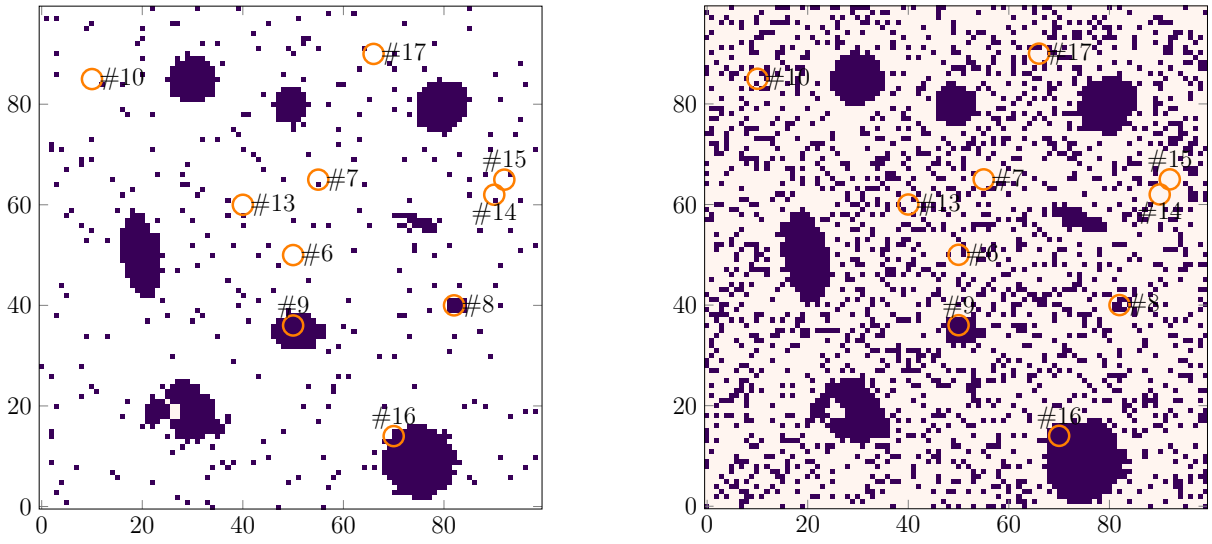
Dans le cas des données MUSE, nous avons modélisé le cube comme la somme des contributions des galaxies et d'un bruit blanc gaussien (dans les trois directions du cube). La détection des raies Lyα dans les spectres du cube constitue un exemple parfait d'application du critère HC^*. Considérons un spectre $\boldsymbol{x} = \mathbf{y}_r(\cdot)$ du cube :

— les échantillons à tester sont les $N = \Lambda = 3600$ valeur d'un spectre,

— sous l'hypothèse $\mathcal{H}_0^i$, les échantillons $\mathbf{y}_r(i)$ d'un spectre sont distribués selon $\mathcal{N}(0, 1)$,

— la fraction $\boldsymbol{\epsilon}$ des échantillons distribués sous $\mathcal{H}_1^i$ correspond aux échantillons de la raie d'émission : $\mathbf{y}_r(i) \sim \mathcal{N}(\mu_i, 1)$, avec la moyenne μ_i non nulle.

La largeur typique d'une raie Lyα étant d'environ 10 bandes spectrales avec la résolution de MUSE pour 3600 bandes observées, la proportion $\boldsymbol{\epsilon}$ de variables pour lesquelles l'hypothèse nulle est rejetée est d'environ $\boldsymbol{\epsilon} \simeq \frac{10}{3600}$ ce qui correspond à $\beta = 0.72$, et à une région de détectabilité définie par $r > \rho^*(\beta) = 0.22$ si $\mu_i \geqslant 1.25$. Il faut donc qu'en moyenne l'amplitude de la raie d'émission soit supérieure à l'écart-type du bruit ($\sigma_\lambda = 1$ pour toutes les longueurs d'onde λ après réduction des données). Les critères HC^*_{3600} et HC^+_{3600} sont appliqués aux données DryRun, centré-réduit par les estimations de moyenne et de variance obtenues par σ-clipping par point fixe, le résultat est présenté sur la figure 3.8. Dans ce cas, le filtrage adapté n'est pas appliqué afin de préserver l'indépendance des échantillons d'un spectre sous l'hypothèse nulle. La probabilité de fausse alarme a été fixée à $\alpha = 5\%$ pour définir le seuil $h(N, \alpha)$ permettant de décider entre l'union des hypothèses nulles ou un élément de son complément. Sur la figure 3.8, un certain nombre de pixels isolés ont été classés dans $\mathcal{C}_1$ alors qu'il s'agit de spectre ne contenant que du bruit, et les galaxies de type Lyα lointaines ne sont détectées précisément (position du centre) par aucune des deux méthodes.



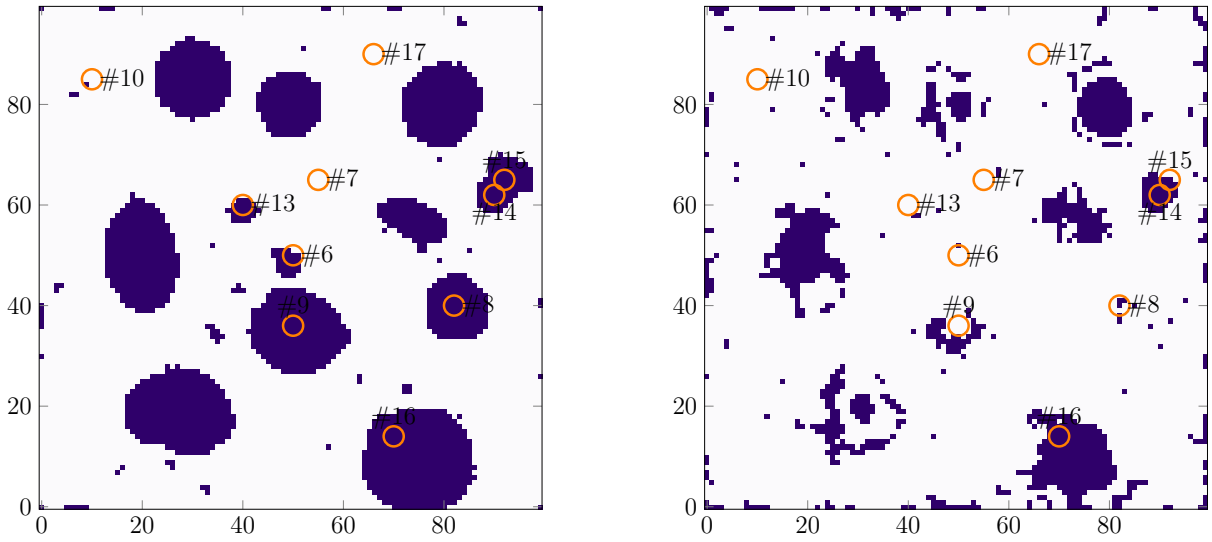
(a) Application du critère HC^*_{3600} (b) Application du critère HC^+_{3600}

FIGURE 3.8 – Résultats du seuillage du cube DryRun, centré et réduit par σ-clipping par point fixe, par les critères HC^*_{3600} (a) et HC^+_{3600} (b). Les cercles oranges modélisent les galaxies de type Lyα dans les données DryRun.

Sans filtrage adapté, les galaxies de type Lyα ne présentent pas une moyenne $\mu_i \geqslant 1.25$ sur toute la largeur de la raie. Appliquer le critère du HC^* ou sa version HC^+ pour seuiller les données sans les prétraiter à l'aide du filtrage adapté ne permettra pas de détecter les objets de très faible intensité avec un spectre contenant une proportion $\boldsymbol{\epsilon} \simeq \frac{10}{3600}$ d'échantillons de moyenne non nulle.

Lorsque nous appliquons le filtrage adapté à la PSF spectralement élargie, les spectres filtrés $\boldsymbol{x} = \mathbf{y}_r^{(f)}$ du bruit sont alors distribués selon une loi gaussienne multivariée de matrice de covariance $\Sigma = C$ dont la définition est donnée dans le paragraphe 3.3.2 : $\mathbf{y}_r^{(f)} \sim \mathcal{N}(\mathbf{0}, C)$. Si le iHC^* est asymptotiquement plus puissant que le HC^* sur des données corrélées, en pratique, les résultats obtenus sur le DryRun pour un même niveau de contrôle sont moins bons en terme de détection et d'erreur de détection. Ceci est illustré sur la figure 3.9 où les critères HC^* et le

iHC sont appliqués aux données après filtrage adapté. La procédure du HC^* appliquée à des tests corrélés (par la LSF dans l'étape de filtrage adapté) ne garantit plus le contrôle optimal asymptotique, cependant les résultats sont bien plus satisfaisant que l'application du HC^* sur les données sans filtrage adapté. Le résultat de l'application du critère HC^+ ne donne en revanche pas un bon résultat en terme de détections et d'erreurs de détection sur les données après filtrage adapté.



(a) Application du critère HC^*_{3600}

(b) Application du critère iHC_{3600}

FIGURE 3.9 – Résultats du seuillage du cube DryRun, centré et réduit, après filtrage adapté, par σ-clipping par point fixe, par les critères HC^*_{3600} (a) et iHC_{3600} (b) avec un contrôle de la probabilité de fausse alarme $\alpha = 1\%$. Les cercles oranges modélisent les galaxies de type Lyα dans les données DryRun.

Donoho and Jin [2004] et Hall et al. [2010] ont montré que toutes ces procédures, HC^*, HC^+ et iHC, sont asymptotiquement ($N \to \infty$) plus puissante pour la détection d'événements rares dans une série d'échantillon. En pratique, sur des échantillons de taille finie ($N \simeq 3600$ pour les spectres des données MUSE), ces méthodes ne sont pas suffisamment performantes. Seule l'application du HC^* sur les données après filtrage adapté présente un intérêt. Ces derniers résultats pourront être comparés à ceux obtenu à l'aide du test proposé dans le paragraphe 3.3.

3.2.2 Seuillage des données par contrôle du FDR

Le contrôle du taux de fausses découvertes sera appliqué dans notre cas directement sur le cube de données après filtrage adapté. Le seuillage par différentes procédures de contrôle du FDR sera donc un cube de données binaires, où chaque pixel pour lequel l'hypothèse nulle $\mathcal{H}_0$ aura été rejetée prendra la valeur 1, et 0 sinon. Quelques rares procédures permettant de contrôler le FDR dans certains cas de dépendance ont été proposées dans la littérature. Récemment dans les travaux de Barber and Candes [2014] les auteurs proposent une méthode de sélection de variables, appelée le *knockoff filter*, qui contrôle le FDR dans le cas du modèle de mélange linéaire. Dans les travaux de Benjamini and Yekutieli [2001], un facteur correctif est ajouté dans la procédure BH afin de conserver le contrôle quelle que soit la structure de dépendance ; cette méthode s'avère bien trop conservative pour être utilisée. Les auteurs montrent également qu'il est possible de contrôler le FDR dans le cas de dépendance avec la procédure originale BH, qui est moins conservative, sous certaines conditions de positivité. Dans le paragraphe 3.4, nous nous

appuierons sur les résultats de Benjamini and Yekutieli [2001] pour montrer qu'il est possible de contrôler le FDR en seuillant le résultat d'un filtrage adapté avec la procédure BH sous réserve de satisfaire quelques conditions réalistes dans le cadre de la détection de sources.

3.2.2.1 Contrôle du FDR par procédure BH dans le cas de tests dépendants

Dans les travaux de Benjamini and Yekutieli [2001], les auteurs proposent d'étendre la procédure de contrôle du FDR proposée par Benjamini and Hochberg [1995], initialement développée pour des tests indépendants, au cas de la dépendance des tests. Deux aspects sont présentés :

- l'utilisation de la procédure originale de Benjamini and Hochberg [1995] sous réserve que la corrélation des tests respecte certaines conditions,
- la correction de la procédure de contrôle du FDR proposée par Benjamini and Hochberg [1995] pour prendre en compte tout type de corrélation entre les tests.

Le résultat le plus souvent utilisé en traitement d'image (voir Hopkins et al. [2002], Genovese et al. [2002] et Whiting [2012]) est la procédure de contrôle du FDR de Benjamini and Hochberg [1995] corrigée afin de prendre en compte les corrélations. Pourtant la prise en compte de la corrélation directement dans l'étape de recherche du seuil conduit à une procédure trop conservative pour être utile afin de rechercher des signaux de faible intensité. Cette procédure corrigée est détaillée dans l'encadré 3.5. Pour tout type de corrélation, la correction consiste à modifier le seuil qui intervient dans le critère par un facteur multiplicatif $c = \sum_{i=1}^{N} \frac{1}{i}$. Ce facteur prend implicitement en considération le fait que l'ensemble des tests sont corrélés. Afin d'obtenir une procédure un peu moins conservative, dans le cas de corrélation locale des tests, Hopkins et al. [2002] envisagent de restreindre la somme $\sum_{i=1}^{N_l} \frac{1}{i}$, où N_l est le support, en pixels, des corrélations locales (typiquement la taille de la PSF). Cependant cette méthode n'offre aucun cadre théorique de contrôle et reste trop conservative en pratique.

ENCADRÉ 3.5 – Procédure de Benjamini-Hochberg corrigée

1. Soit $\mathcal{H}_0^{(1)}$, ..., $\mathcal{H}_0^{(N)}$ une famille d'hypothèses nulles et $p_1, ..., p_N$ les p-valeurs correspondantes.
2. Notons $p_{(0)} < p_{(1)} \leqslant ... \leqslant p_{(N)}$ les p-valeurs ordonnées de façon croissante et $\mathcal{H}_0^{(1)}$, ..., $\mathcal{H}_0^{(N)}$ les hypothèses nulles associées. Par convention $p_{(0)} = 0$.
3. Soit $k = \underset{i}{\text{argmax}} \left(p_{(i)} \leqslant q \frac{i}{cN} \right)$ avec $c = \sum_{i=1}^{N} \frac{1}{i}$
4. Rejet des hypothèses nulles $\mathcal{H}_0^{(1)}$, ..., $\mathcal{H}_0^{(k)}$ et acceptation des hypothèses $\mathcal{H}_0^{(k+1)}$, ..., $\mathcal{H}_0^{(N)}$.

Dans le cas des données MUSE, utiliser le facteur correctif pour contrôler le FDR mènerait à une procédure tellement conservative que seuls les objets de très forte intensité seraient détectés, dans ce cas, $N = 324$ millions de pixels. Même en restreignant le facteur correctif au support de la PSF de l'instrument (qui corrèle les tests lors de l'étape de filtrage adapté), $N = 3087$ ce qui rend la procédure encore bien trop conservative.

3.2.2.2 Knockoff filter

Barber and Candes [2014] se sont intéressés au problème du contrôle du FDR dans les procédures de sélection de variables pour un modèle de mélange linéaire gaussien :

$$\boldsymbol{y} = \mathbf{X}\beta + \boldsymbol{z},$$

où $\boldsymbol{y} \in \mathbb{R}^n$ est le vecteur des observations, $\mathbf{X} = [\mathbf{X}_1, \cdots, \mathbf{X}_p] \in \mathbb{R}^{n \times p}$ est une matrice de design connue, $\beta \in \mathbb{R}^p$ est le vecteur de coefficients à estimer et $\boldsymbol{z} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$ est un bruit gaussien. L'intérêt de leur méthode est de prendre en compte explicitement la structure de corrélation afin d'obtenir une procédure puissante qui continue de contrôler le FDR. Le détail de la procédure est donné dans l'annexe G.

Cette méthode présente, *a priori*, un grand intérêt pour notre problème puisque l'on se trouve exactement dans le cas d'un modèle de mélange linéaire gaussien avec des variables corrélées par le filtrage adapté. La formulation du problème de seuillage par contrôle du FDR que nous présenterons dans le paragraphe 3.4 est similaire à celle introduite dans le papier de Barber and Candes [2014]. Les $\beta_j = (\mathbf{H}\boldsymbol{a})_j$ correspondent aux contributions des sources $\boldsymbol{a}$ élargies par la réponse impulsionnelle $\mathbf{H}$ de l'instrument dans le j^{eme} pixel, et nous souhaitons dresser une liste $\hat{S}$ de pixels contenant probablement la contribution d'une ou plusieurs sources tout en contrôlant le taux de fausses découvertes dans cette liste. La matrice de design $\mathbf{X}$ correspond dans notre cas à la matrice $\mathbf{H}$ modélisant l'opération de filtrage adapté à la PSF de l'instrument, chaque colonne de $\mathbf{H}$ contient la PSF centrée en (p, q, λ) pour tout $(p, q, \lambda) \in [0, P] \times [0, Q] \times [0, \Lambda]$. L'utilisation du *knockoff filter* est cependant impossible à cause de différents points :

- la dimension des données, pour nos données, $n = p \simeq 3 \times 10^8$, qui est rédhibitoire pour la construction des contrefaçons $\tilde{\mathbf{X}}$ nécessaires à la procédure, *cf.* annexe G,
- le fait que l'on ait autant de variables à tester que d'observations ajoute également un obstacle à l'utilisation du *knockoff filter* : nous nous retrouvons dans le cas particulier $n = p$ où il faut augmenter artificiellement le vecteur d'observations de façon à obtenir $n \geqslant 2p \simeq 6 \times 10^8$ échantillons,
- les corrélations induites par le filtrage adapté sont très localisées (la taille de la PSF est faible devant la taille des données) mais elles sont très fortes et très structurées. Même sur des jeux de données de taille très réduite, la construction de confre-façons qui respectent les contraintes sur les corrélations formulées par les équations (G.2), (G.3) et (G.4) est compliquée : ces contrefaçons ressemblent trop aux originaux et la puissance de la procédure s'effondre (nous obtenons un détecteur aléatoire).

3.3 Contrôle du FWER : un premier prétraitement basé sur le filtrage adapté et la statistique du maximum des spectres

Une fois la détectabilité des objets améliorée par le filtrage adapté (PSF spectralement élargie (d)), nous devons définir un test pour effectuer le seuillage des données. Nous nous intéressons ici à un test d'hypothèses définie pour chaque spectre filtré $\mathbf{y}_r^f(\cdot)$, à la position spatiale r, que nous définissons de la façon suivante :

$$\begin{cases} \mathcal{H}_0 & : \text{ le spectre contient uniquement du bruit} \\ \mathcal{H}_1 & : \text{ il existe au moins une raie d'émission d'une source + du bruit} \end{cases} \tag{3.9}$$

Dans le paragraphe 3.1.4.2, nous avons utilisé, pour afficher les performances des différents filtrages, la projection définie par l'équation (3.3). Nous formulons maintenant cette projection comme un test T sur chaque spectre $\mathbf{y}_r^f(\cdot)$ du cube à la position $(p, q) \equiv r$:

$$T(\mathbf{y}_r^f(\cdot)) = \max_{\lambda}(\mathbf{y}_r^f(\lambda)) \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\lessgtr}} \eta, \tag{3.10}$$

Fixer le seuil η demande l'écriture de la statistique du test T sous l'hypothèse nulle et définir le critère d'erreur que nous souhaitons contrôler.

Le test T (voir équation (3.10)) compare la valeur maximale d'un spectre après filtrage adapté à une valeur de seuil η pour décider si ce spectre contient la contribution d'une source ou non. Nous allons nous intéresser en détail à ce test que nous appellerons max-test par la suite, en référence aux travaux de Arias-Castro et al. [2011] qui s'intéressent également à la statistique du maximum d'un ensemble de test. Notons que ce test, bien qu'obtenu par une philosophie différente (filtrage adapté et contrôle d'événements rares), est similaire au test 1.9 formulé par Paris et al. [2013b] (test de vraisemblance généralisé sous contrainte).

3.3.1 Principe général

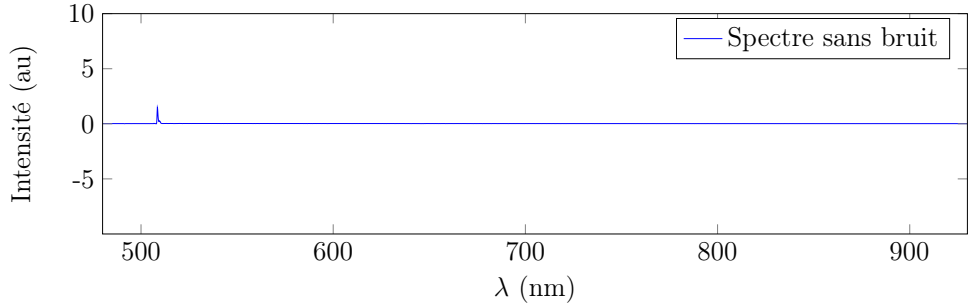
Les galaxies que nous cherchons à détecter ont une réponse en trois dimensions qui ressemble à la PSF de l'instrument MUSE. Le filtrage adapté doit améliorer la détectabilité de ce type de sources en maximisant leur rapport signal à bruit, notamment au niveau de la raie d'émission Lyα comme illustré sur la figure 3.10. Alors que la raie d'émission est totalement invisible dans le spectre bruité, après filtrage adapté des données, on retrouve la caractéristique de la raie Lyα au milieu du bruit, qui est fortement atténué par le filtrage.

Une statistique suffisante pour détecter ce type de galaxies est donc de se concentrer sur la valeur maximale du spectre après filtrage adapté, ce qui conduit à un contrôle de type FWER. Le contrôle du FWER revient à contrôler la probabilité qu'au moins une hypothèse $\mathcal{H}_0$ soit rejetée à tort, *i.e.* qu'au moins une valeur du spectre est suffisamment grande. Contrôler si au moins la valeur maximale du spectre est supérieure au seuil de décision η dans le test défini par l'équation 3.10, revient à contrôler le critère du FWER pour le cas défini par (3.9). En effet, le critère du FWER s'écrit :

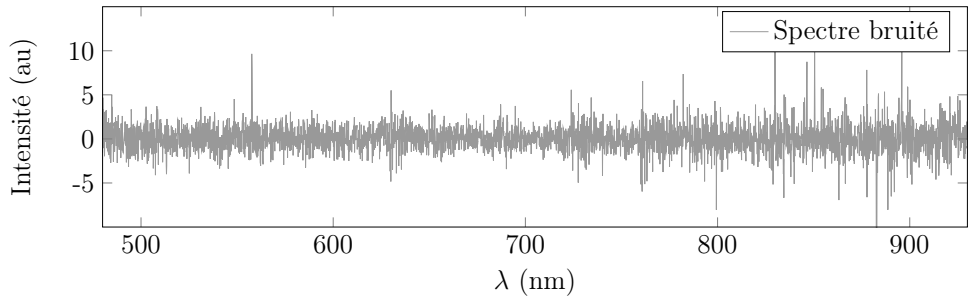
$$\begin{aligned} FWER &= Pr\left(\exists \lambda \in \Lambda_0 \text{ t.q. } \mathbf{y}_r^f(\cdot) > \eta\right) \\ &= Pr\left(\max_{\lambda\in\Lambda_0}\left(\mathbf{y}_r^f(\cdot)\right) > \eta\right) \\ &\leqslant Pr\left(\max_{\lambda}\left(\mathbf{y}_r^f(\cdot)\right) > \eta|\mathcal{H}_0\right) \end{aligned} \tag{3.11}$$

En comparant la valeur maximale de chaque spectre après filtrage adapté, nous obtenons une carte de détection binaire avec la classe $\mathcal{C}_0$ des pixels ne contenant probablement que du bruit (la valeur maximale des spectres de cette classes étant inférieure à la valeur η du seuil) et la classe $\mathcal{C}_1$ des pixels appartenant avec une certaine probabilité d'erreur à une source.

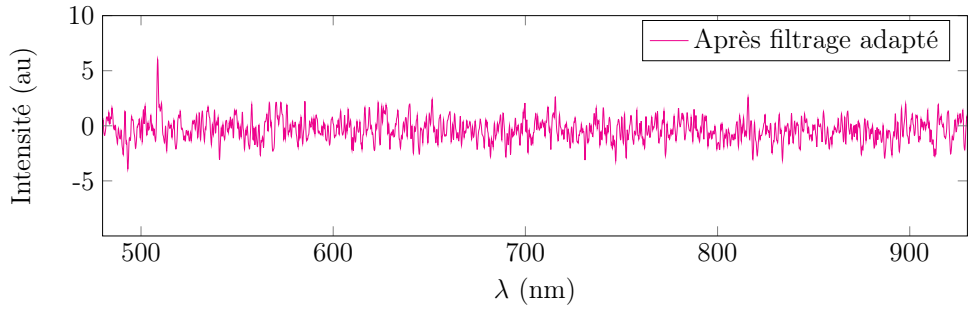
Le principe général de la méthode basée sur l'application du max-test aux spectres en sortie du filtrage adapté est résumé dans l'encadré 3.6. Nous allons ensuite nous intéresser aux propriétés de ce test, et notamment à la statistique du max-test dans le cas d'échantillons gaussiens multivariés.



(a) Données originales



(b) Données centrées-réduites



(c) Données originales

FIGURE 3.10 – Représentation du spectre du centre d'une galaxie de type Lyα de faible intensité : (a) sans bruit, (b) avec bruit additif gaussien et (c) après filtrage adapté.

ENCADRÉ 3.6 – Max-test appliqué à la sortie du filtrage adapté

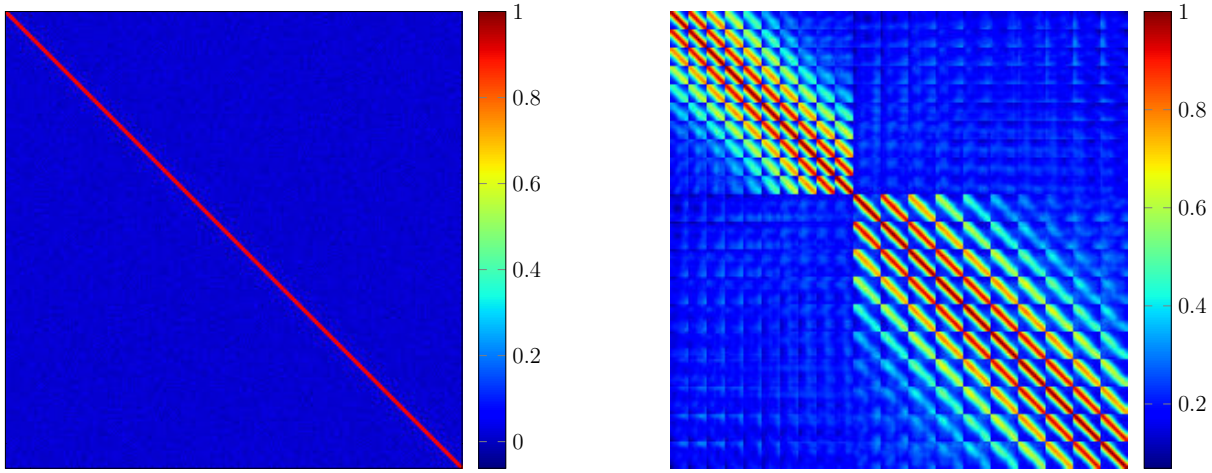
1. Filtrage adapté : $\mathbf{Y}^{f,(v)} = \mathbf{H}^T \mathbf{Y}^{(v)}$
2. Choix d'un seuil η en fonction de la probabilité de fausse alarme, détermination à partir de la statistique du test T sous l'hypothèse nulle.
3. Max-test appliqué à chaque spectre $\mathbf{y}_r^f = \mathbf{Y}^f(p, q, \cdot)$ à la position $r \equiv (p, q)$:

$$T(\mathbf{y}_r^f(\cdot)) = \max_{\lambda}(\mathbf{y}_r^f(\lambda)) \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\lessgtr}} \eta,$$

4. Construction de la carte de proposition en fonction du résultat de T en chaque position spatiale (p, q) : le pixel (p, q) vaut 1 si $T(\mathbf{y}_r^f(\cdot)) > \eta$ et 0 sinon.

3.3.2 Filtrage adapté

Le filtrage adapté à la PSF de l'instrument est appliqué au cube centré-réduit calculé précédemment. Si, sous l'hypothèse nulle $\mathcal{H}_0$, tous les pixels sont supposés indépendants spectralement et spatialement, le filtrage adapté introduit une corrélation entre les pixels voisins (dans les trois dimensions) et donc également entre les spectres voisins. Nous représentons sur la figure 3.11 la matrice de corrélation (spatiale) empirique S des 250 spectres extraits de deux zones de bruit du cube DryRun, avant (figure 3.11a) et après (figure 3.11b) avoir appliqué le filtrage adapté. Cette matrice S représente la composante spatiale de la corrélation induite par le filtrage adapté. Nous ne nous intéresserons pas à cette corrélation spatiale pour le max-test puisque chaque spectre est testé séparément.



(a) Avant filtrage adapté (b) Après filtrage adapté

FIGURE 3.11 – Matrice S des corrélations empiriques spatiales des 250 spectres extraits de deux zones de bruit du cube DryRun, avant filtrage adapté (a) et après filtrage adapté (b). Avant filtrage adapté, la matrice est proche de l'identité, ce qui montre bien l'indépendance spatiale des données sous l'hypothèse de bruit seul, tandis qu'on retrouve une structure de corrélation après le filtrage adapté. La structure par bloc de la matrice de corrélation après filtrage adapté est dû à l'arrangement des spectres utilisés pour calculer la covariance (un bloc modélise une zone, et le quadrillage à l'intérieur des blocs est dû au l'arrangment des spectres en lignes et en colonnes dans le cube).

On se propose ici de définir la corrélation spectrale induite par le filtrage adapté sur les pixels d'un spectre du cube de données sous l'hypothèse $\mathcal{H}_0$. Soit $\mathbf{y}_r^f$ le spectre correspondant à la position $r \equiv (p, q)$ du cube distribué selon $\mathbf{y}_r^f \sim \mathcal{N}(\mathbf{0}, C)$, où $\mathbf{0}$ est le vecteur nul de taille $\Lambda \times 1$ et C est la matrice de covariance. La matrice de covariance C est symétrique, creuse, et ses éléments C_{λ_1,λ_2} s'écrivent :

$$\begin{aligned} C_{\lambda_1,\lambda_2} &= \operatorname{cov}\left(\mathbf{y}_r^f(\lambda_1), \mathbf{y}_r^f(\lambda_2)\right) \\ &= \operatorname{cov}\left(\sum_{\mu_1} \mathbf{Z}_{\mu_1} L_{\mu_1}(\lambda_1) \sum_{\mu_2} \mathbf{Z}_{\mu_2} L_{\mu_2}(\lambda_2)\right) \\ &= \sum_{\mu_1}\sum_{\mu_2} L_{\mu_1}(\lambda_1) L_{\mu_2}(\lambda_2)\ \operatorname{cov}\left(\mathbf{Z}_{\mu_1}, \mathbf{Z}_{\mu_2}\right) \end{aligned}$$

où $\mathbf{Z}_{\mu_i} = \sum_z \mathbf{y}_z(\mu_i) F_{\mu_i}(r - z)$ est la convolution des données par la FSF de MUSE effectuée dans

le filtrage adapté. Le terme $\mathrm{cov}\,(\mathbf{Z}_{\mu_1}, \mathbf{Z}_{\mu_2})$ peut s'écrire :

$$\mathrm{cov}\,(\mathbf{Z}_{\mu_1}, \mathbf{Z}_{\mu_2}) = \sum_{z_1} \sum_{z_2} F_{\mu_1}(z_1) F_{\mu_2}(z_2)\ \mathrm{cov}\,(\mathbf{y}_{z_1}(\mu_1), \mathbf{y}_{z_2}(\mu_2))$$

et sous l'hypothèse nulle, $\mathrm{cov}\,(\mathbf{y}_{z_1}(\mu_1), \mathbf{y}_{z_2}(\mu_2)) \neq 0$ si et seulement si les deux conditions suivantes sont vérifiées :

$$\begin{cases} A_1 & : & \mu_1 & = & \mu_2 \\ A_2 & : & z_1 & = & z_2 \end{cases}$$

alors $\mathrm{cov}\,(\mathbf{y}_{z_1}(\mu_1), \mathbf{y}_{z_2}(\mu_2)) = \sigma^2_{\mu_1}$. Finalement :

$$\begin{cases} \mathrm{cov}\,(\mathbf{Z}_{\mu_1}, \mathbf{Z}_{\mu_2}) & = & \sigma^2_\mu \sum\limits_z F_\mu(z)^2 & \text{si } A_1 \text{ et } A_2, \\ \mathrm{cov}\,(\mathbf{Z}_{\mu_1}, \mathbf{Z}_{\mu_2}) & = & 0 & \text{sinon.} \end{cases}$$

La matrice de covariance C est symétrique, considérons le cas $\lambda_1 \leqslant \lambda_2$, l'élément C_{λ_1,λ_2} de la matrice C s'écrit :

$$\begin{cases} C_{\lambda_1,\lambda_2} & = & \sum\limits_{\mu=\lambda_2-\frac{\Delta_{LSF}}{2}}^{\lambda_1+\frac{\Delta_{LSF}}{2}} L_\mu(\lambda_1) L_\mu(\lambda_2) \sigma^2_\mu \sum\limits_z F_\mu(z)^2 & \text{si } \left|\lambda_1 - \lambda_2\right| \leqslant \Delta_{LSF}, \\ C_{\lambda_1,\lambda_2} & = & 0 & \text{sinon.} \end{cases}$$

Le cas $\lambda_1 \geqslant \lambda_2$ est obtenu en inversant les rôles de λ_1 et λ_2 dans le système ci-dessus. Puisque la FSF est normalisée (norme ℓ_2), $\sum\limits_z F_\mu(z)^2 = 1$ et les données sont réduites, $\sigma^2_\mu = 1$, l'expression de la covariance peut se simplifier :

$$\begin{cases} C_{\lambda_1,\lambda_2} & = & \sum\limits_{\mu=\lambda_2-\frac{\Delta_{LSF}}{2}}^{\lambda_1+\frac{\Delta_{LSF}}{2}} L_\mu(\lambda_1) L_\mu(\lambda_2) & \text{si } \left|\lambda_1 - \lambda_2\right| \leqslant \Delta_{LSF}, \\ C_{\lambda_1,\lambda_2} & = & 0 & \text{sinon.} \end{cases} \tag{3.12}$$

La figure 3.12 représente graphiquement la matrice de covariance théorique C définie par l'équation 3.12 pour les 25 premières longueurs d'onde.

3.3.3 Apprentissage de la loi du test sous $\mathcal{H}_0$

La statistique de test définie par l'équation 3.10 consiste à prendre la valeur maximale d'un vecteur pour lequel nous allons faire plusieurs hypothèses quant à sa distribution. La loi d'une telle statistique de test est difficilement exprimable analytiquement, nous allons nous intéresser à différents moyens d'estimer cette loi.

3.3.3.1 Loi théorique dans le cas d'un bruit i.i.d.

Dans ce paragraphe, nous supposons que le bruit présent sur l'image est i.i.d. de loi $\mathcal{N}(0,1)$, et donc que la covariance des éléments d'un spectre après filtrage adapté est la matrice C définie par 3.12. Il n'est pas possible d'exprimer analytiquement la statistique de la valeur maximale

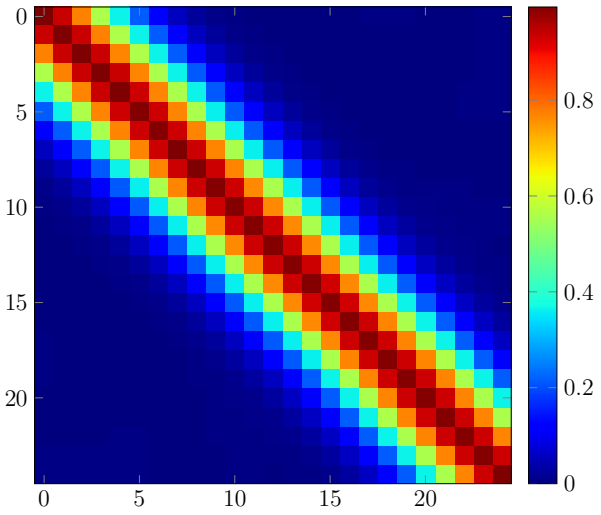


FIGURE 3.12 – Représentation graphique de la matrice de covariance théorique C pour les 25 premières longueurs d'onde.

d'une vecteur gaussien multivarié, il faudra donc calculer numériquement à l'aide de méthode de Monte Carlo la loi de cette statistique. Une solution consiste à générer un grand nombre de vecteurs $\boldsymbol{x}$ tels que $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_\Lambda)$ qui sont transformés en vecteurs gaussiens multivariés $\tilde{\boldsymbol{x}}$ de loi $\mathcal{N}(\mathbf{0}, C)$ par l'opération :

$$\tilde{\boldsymbol{x}} = B\boldsymbol{x},$$

avec B la matrice triangulaire inférieure de la décomposition de Cholesky de la matrice de covariance $C = BB^T$. Pour estimer la loi de la statistique du test T sous l'hypothèse nulle, il suffit de calculer T (équation 3.10) sur les vecteurs $\tilde{\boldsymbol{x}}$ générés et d'utiliser la distribution empirique du test représentée sur la figure 3.13. Puisque la distribution du bruit est symétrique et que les données sont centrées, la loi du max-test doit être la même que la loi du min-test (au signe près) [5], ce qui se vérifie sur la figure 3.13. A partir de cette distribution empirique, il est possible de choisir une valeur de seuil η en fonction de la probabilité de fausses alarmes p_{FA} désirée. Soit $\mathbf{y}_r^f$ un spectre après filtrage adapté, et $T = \max_\lambda(\mathbf{y}_r^f(\lambda))$, le max-test qui consiste à tester la valeur maximale du spectre $\mathbf{y}_r^f$. Décider de classer $\mathbf{y}_r^f$ sous l'hypothèse $\mathcal{H}_0$ si $T \leqslant \eta = 4.05$ et sous $\mathcal{H}_1$ sinon, revient à accepter une probabilité de fausses alarmes de 5% d'après la courbe bleue de la figure 3.13.

3.3.3.2 Apprentissage non paramétrique de la loi à partir des valeurs minimales des spectres

Nous cherchons à estimer la distribution de la statistique du max-test T sous $\mathcal{H}_0$. Cette loi ne peut pas être apprise à partir de la loi empirique de T dans les données observées. En effet la loi empirique de T va évidemment être biaisée par la présence des sources dans l'image. En revanche, en supposant que la contribution des sources est toujours positive et que le bruit est symétrique et convenablement centré, le min-test ne devrait pas être impacté par la présence

5. Le min-test désigne le test qui consiste à prendre la valeur minimale d'un vecteur et à changer le signe de cette valeur. Cette opération est identique à celle qui consiste à prendre le max-test de l'opposé du vecteur. Ceci permet d'exploiter le fait que sous $\mathcal{H}_0$, si la distribution du bruit est symétrique, les minima des vecteurs devraient être similaire aux maxima des vecteurs.

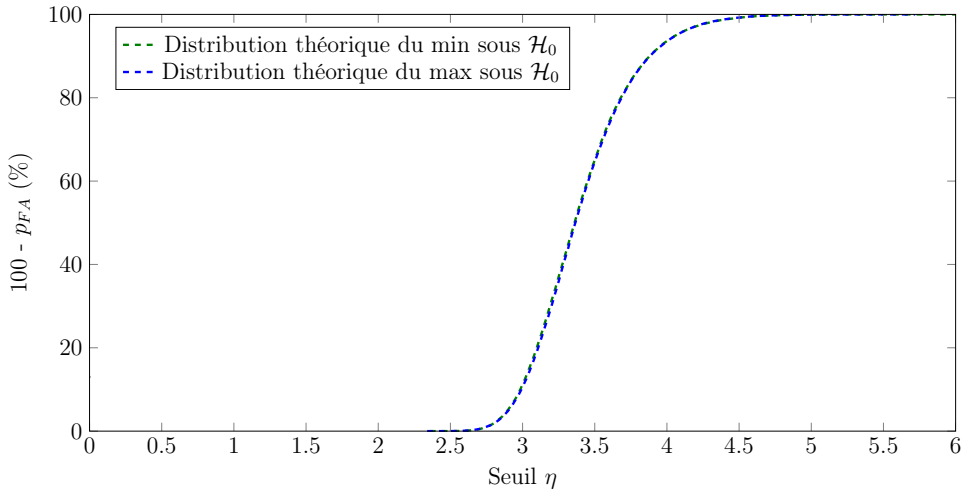


FIGURE 3.13 – Fonctions de répartition empiriques du max-test et du min-test sous l'hypothèse $\mathcal{H}_0$ en considérant qu'avant filtrage adapté les éléments d'un spectre sont i.i.d. de loi $\mathcal{N}(0,1)$. Ces courbes permettent de calculer le seuil de décision η en fonction de la probabilité de fausse alarme p_{FA} choisie.

de sources et l'allure de la loi répartition des minima des spectres devraient être similaire aux courbes obtenues sur la figure 3.13.

Afin d'illustrer la loi de la statistique du minimum des spectres, nous générons un nouveau cube de données avec un certain nombre de sources ($\sim$ 50 sources) dont quelques unes ont un spectre continu de forte intensité et les autres un spectre nul presque partout sauf sur quelques longueurs d'onde consécutives. Spatialement, ces sources ont un profil d'intensité gaussien. Un bruit blanc gaussien i.i.d. est ajouté à ces sources pour former le cube de données, nous utilisons un bruit gaussien centré-réduit i.i.d. afin de contrôler tous les paramètres de la simulation et pouvoir comparer aux résultats obtenus par méthode de Monte Carlo présenté sur la figure 3.13. Le filtrage adapté est ensuite appliqué à ce cube de données. Les courbes obtenues pour le min-test et le max-test sur les données synthétiques sont présentées sur la figure 3.14.

Si la courbe du min-test calculée sur les données synthétiques contenant des sources converge bien vers la distribution empirique du max-test sous $\mathcal{H}_0$ pour les valeurs de seuils supérieures à $\eta > 4.06$ (*i.e.* pour des probabilités de fausses alarme inférieures à 5%), elle diffère pour les petites valeurs de seuils. Ceci s'explique par la présence de sources à spectre continu, dont l'intensité est significative tout au long du spectre. Le min-test appliqué à un tel spectre prendra une valeur stochastiquement plus élevée que pour les spectres générés sous $\mathcal{H}_0$. Ce n'est pas gênant en pratique puisque nous n'utiliserons pas de probabilité de fausses alarmes plus élevées que 5%. **Sous réserve que la distribution du bruit soit bien symétrique, utiliser le min-test calculé sur les données, permet alors de calibrer le seuil à utiliser pour le max-test.**

3.3.3.3 Apprentissage paramétrique : estimation de la matrice de covariance

Sous l'hypothèse de gaussianité du bruit, il suffit d'estimer la matrice de covariance des spectres sous $\mathcal{H}_0$ afin de déterminer la loi théorique de la statistique du maximum des spectres. Nous nous restreindrons ici, à l'estimation des covariances sur des vecteurs de taille $d \times 1$ avec $d = 25$, car nous savons que le filtrage adapté à la PSF spectralement élargie entraîne une corrélation des échantillons d'un spectre sur un support de 19 pixels, nous choisissons $d = 25$ pour assurer une marge de sécurité. La matrice de covariance empirique des spectres est obtenue

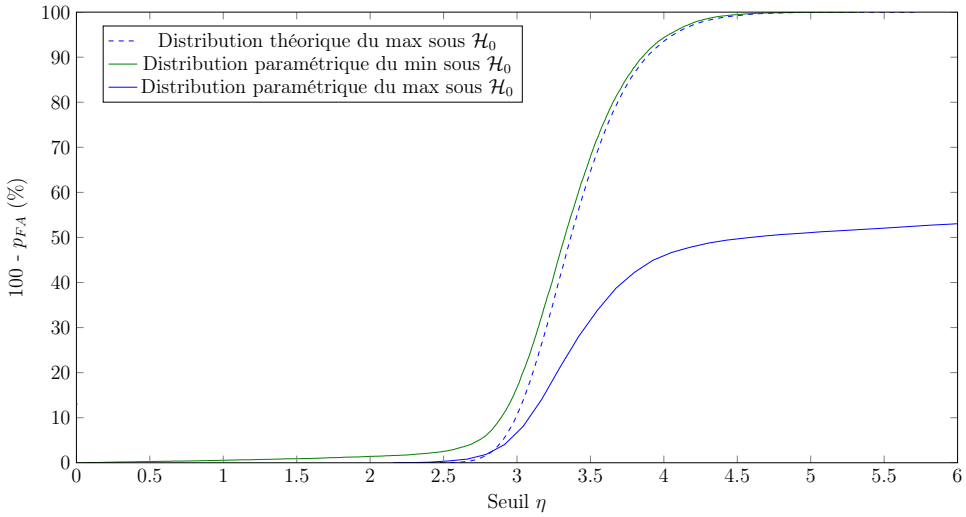


FIGURE 3.14 – Répartition empirique des valeurs maximales et des valeurs minimales calculées sur les spectres d'un cube de données synthétiques contenant des sources. La répartition théorique du max-test sous l'hypothèse nulle est également reportée (courbe en pointillés).

à partir des N spectres x_i extraits des données,

$$\hat{C}_\eta = \frac{1}{N-1}(X_\eta - \mu)(X_\eta - \mu)^T \tag{3.13}$$

où $X_\eta = [x_1, \cdots, x_N]$ est la matrice contenant dans ses colonnes les N vecteurs x_i de taille $d \times 1$ extraits à partir des données seuillées ($\leqslant \eta$) et μ est le vecteur de taille $d \times 1$ correspondant à la moyenne des N vecteurs x_i. Les x_i peuvent être vus comme N réalisations d'un vecteur aléatoire X de taille $d \times 1$. Les données étant centrées, l'expression (3.13) est l'estimateur du maximum de vraisemblance de la matrice de covariance débiaisée par le facteur $\frac{N}{N-1}$.

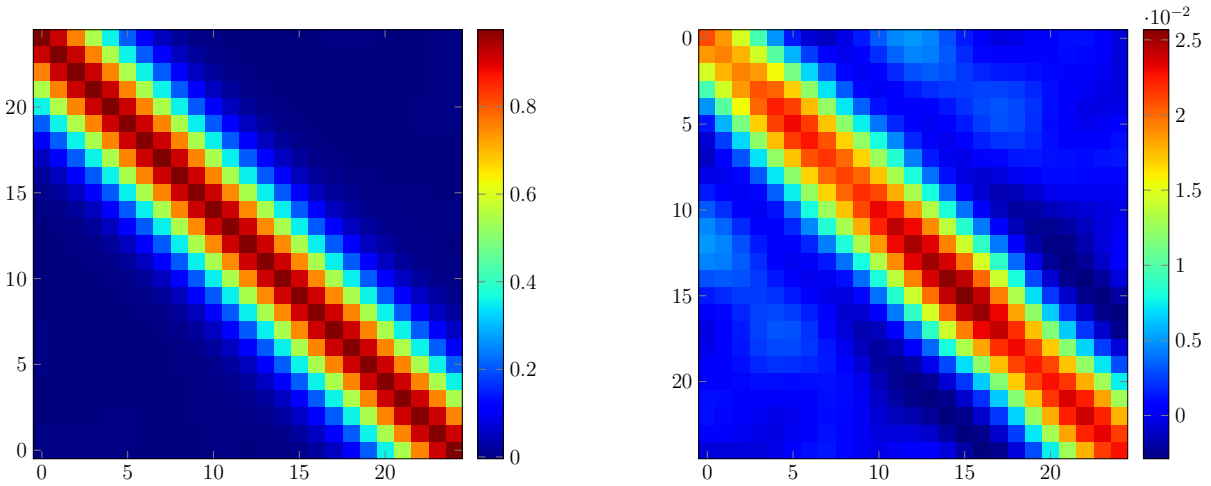


FIGURE 3.15 – Estimation de la matrice de covariance empirique $\hat{C}_{\eta=3.6}$ (gauche) et différence avec la matrice de covariance théorique (droite). Les vecteurs de bruit ont été seuillés à une valeur maximale de $\eta = 3.6$ qui correspond d'après la loi du maximum pour des données i.i.d. sous $\mathcal{H}_0$ à un seuillage très strict du bruit (probabilité de fausses alarmes de 20%).

Les résultats de la figure 3.15 montre que l'estimation de la matrice de covariance d'un bruit i.i.d. filtré est très proche de la matrice de covariance théorique définie par l'équation 3.12. La matrice de covariance empirique est légèrement sous-estimée par rapport à la matrice de covariance théorique. Si en revanche les données sont seuillées à une valeur plus élevée, par exmple $\eta = 6$, certains vecteurs contiendront la contribution de sources et la matrice de covariance empirique sera légèrement supérieure à la matrice théorique comme on peut le voir sur la figure 3.16.

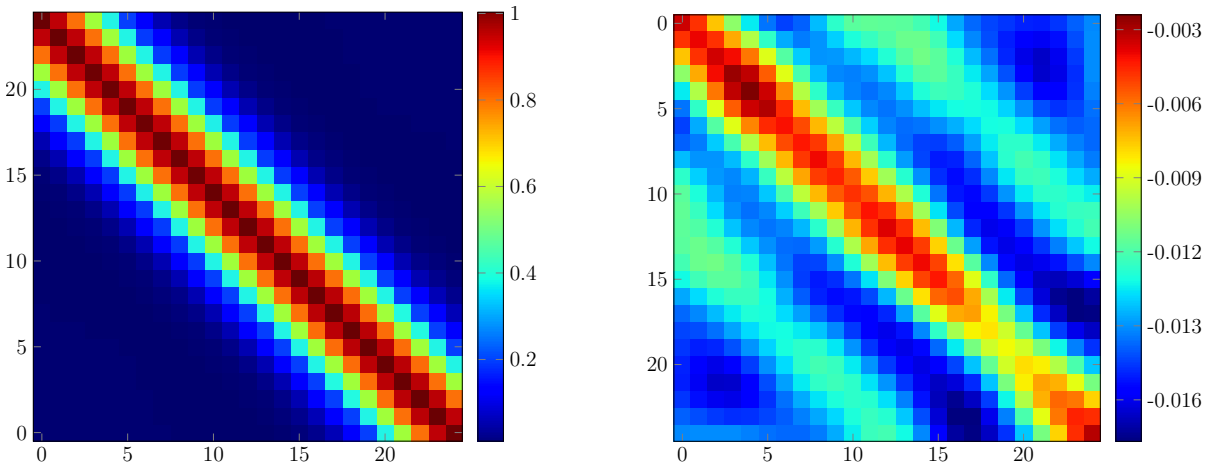


FIGURE 3.16 – Estimation de la matrice de covariance empirique $\hat{C}_{\eta=6}$ (gauche) et différence avec la matrice de covariance théorique (droite). Les vecteurs de bruit ont été seuillés à une valeur maximale de $\eta = 6$ qui correspond d'après la loi du maximum pour des données i.i.d. sous $\mathcal{H}_0$ ne pas perdre de vecteur de bruit, mais aussi à accepter certains spectres appartenant à des sources.

D'après les figures 3.15 et 3.16, le seuillage des données influe peu sur la matrice de covariance empirique des spectres de bruit. Nous souhaitons maintenant étudier l'influence des différents seuillages sur la loi du maximum des spectres i.i.d que nous avons corrélés avec les matrices de covariance apprises. La figure 3.17 présente les distributions obtenues par méthode de Monte Carlo avec les matrices de covariance estimées $\hat{C}_{\eta=3}$, $\hat{C}_{\eta=3.6}$, et $\hat{C}_{\eta=6}$. La troncature a pour effet de décaler la courbe du max-test vers la gauche : se fier à la courbe du max apprise sur les données corrélées par $\hat{C}_{\eta=3}$ entraîne un contrôle beaucoup plus conservatif que pour les courbes obtenues avec $\hat{C}_{\eta=3.6}$, et $\hat{C}_{\eta=6}$. Cependant toutes les courbes convergent de la même façon pour une probabilité de fausses alarmes $p_{FA} \leqslant 5\%$.

3.3.3.4 Apprentissage paramétrique : estimation de la matrice de covariance après centrage des spectres

Dans le paragraphe précédent, la matrice de covariance était estimée à l'aide de N spectres extraits directement des données. L'estimation proposée dans ce paragraphe consiste à apprendre la matrice de covariance en centrant tout d'abord individuellement les N réalisations de X : $x_i^c = x_i - \bar{x}_i \mathbf{1}_d$. L'intérêt de cette approche est de s'affranchir du continu qui peut être présent dans les spectres, et obtenir ainsi une estimation plus précise. La matrice de covariance empirique de la variable $X - \bar{x}\mathbf{1}_d$, où $\bar{x} = \frac{1}{d}\sum_{i=1}^{d} X_i$ est la moyenne du vecteur X, est calculée à partir des N

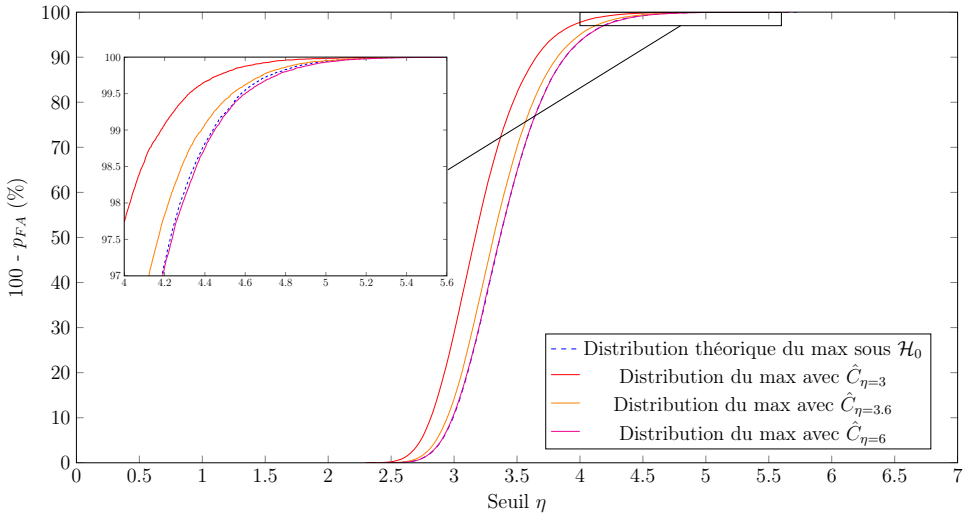


FIGURE 3.17 – Influence de la troncature des spectres sur la loi du max-test obtenue sur les données i.i.d. corrélées par la matrice de covariance estimée (courbes rouge, orange et magenta). Ces courbes sont comparées à la loi empirique du maximum de données i.i.d. (courbe pointillée bleue) obtenue par méthode de Monte Carlo.

vecteurs contenus dans $X_\eta^c = [x_1^c, \cdots, x_N^c]$ par :

$$\begin{aligned}\hat{C}_\eta^c &= \mathrm{cov}(X - \bar{x}\mathbf{1}_d) \\ &= \mathrm{E}\big[(X - \bar{x}\mathbf{1}_d)(X - \bar{x}\mathbf{1}_d)^T\big] \\ &= \frac{1}{N-1} X_\eta^c X_\eta^{cT}\end{aligned}$$

La matrice $\hat{C}_\eta^c$ n'est pas la vraie matrice de covariance des données puisque :

$$\begin{aligned}\hat{C}_\eta^c &= \mathrm{E}\big[\,(X - \mu - (\bar{x}\mathbf{1}_d - \mu))\,(X - \mu - (\bar{x}\mathbf{1}_d - \mu))^T\,\big] \\ &= \mathrm{E}\big[(X-\mu)(X-\mu)^T\big] + \mathrm{E}\big[(\bar{x}\mathbf{1}_d - \mu)(\bar{x}\mathbf{1}_d - \mu)^T\big] \\ &\qquad - \mathrm{E}\big[(X-\mu)(\bar{x}\mathbf{1}_d - \mu)^T\big] - \mathrm{E}\big[(\bar{x}\mathbf{1}_d - \mu)(X-\mu)^T\big] \\ &= \mathrm{cov}(X) + \frac{1}{d^2}\mathbf{1}_d^T\Sigma\mathbf{1}_d\mathbf{U}_d - A \\ &= \Sigma + \frac{1}{d^2}\mathbf{1}_d^T\Sigma\mathbf{1}_d\mathbf{U}_d - A\end{aligned}$$

avec $\mu = \mathrm{E}\big[X\big]$ de taille $d \times 1$, $\Sigma = \mathrm{cov}(X)$ la matrice de covariance des spectres non centrés, A une matrice de taille $d \times d$ telle que ses éléments $A_{i,j} = \frac{1}{d}\mathbf{1}_d\Sigma_{:,i} + \frac{1}{d}\mathbf{1}_d\Sigma_{:,j}$ et $\mathbf{U}_d$ la matrice unité de taille $d \times d$. La matrice $\hat{C}_\eta^c$ n'est pas la matrice de covariance des éléments d'un spectre, il faut corriger cette matrice pour obtenir la matrice de covariance qui nous intéresse. Ceci peut être obtenu numériquement en résolvant l'équation du point fixe suivante :

$$\Sigma = \hat{C}_\eta^c - \frac{1}{d^2}\mathbf{1}_d^T\Sigma\mathbf{1}_d\mathbf{U}_d + A \tag{3.14}$$

L'estimation de Σ ainsi obtenue sera notée : $\hat{C}_\eta^{corr}$ sur les figures 3.18 et 3.20.

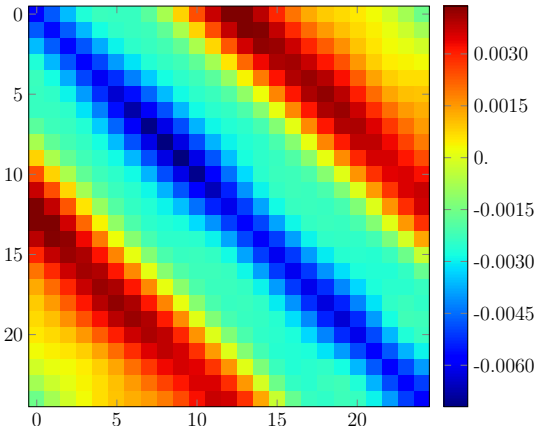


FIGURE 3.18 – Différence entre la matrice de covariance théorique et la matrice de covariance corrigée $\hat{C}_{\eta}^{corr}$, estimée à partir des vecteurs centrés.

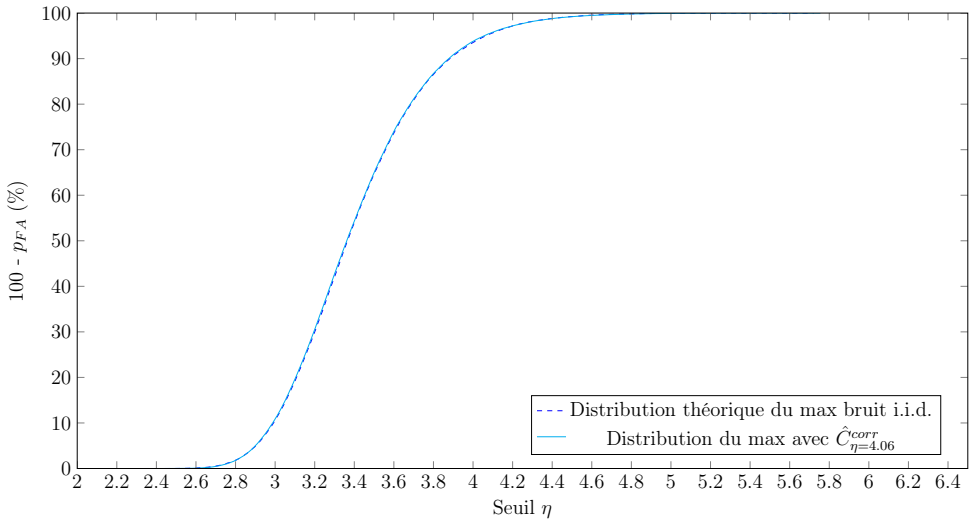


FIGURE 3.19 – Fonction de répartition de la valeur maximale obtenue par méthode de Monte Carlo sur des données i.i.d. corrélées par la matrice de covariance estimée. Cette courbe est comparée à la loi théorique du maximum de données i.i.d. (courbe pointillée bleue) obtenue par méthode de Monte Carlo.

3.3.3.5 Bilan sur l'apprentissage de la loi des valeurs maximales sous $\mathcal{H}_0$

Nous présentons ici un résumé des différentes méthodes étudiées pour apprendre la loi du max-test sous l'hypothèse $\mathcal{H}_0$. Nous avons vu que si la structure de covariance est correctement estimée à partir de vecteur de bruit extrait des données, il est possible d'estimer la distribution théorique de la valeur maximale des spectres sous l'hypothèse paramétrique gaussienne.

Plusieurs méthodes ont été étudiées pour apprendre la structure de covariance, elles donnent des résultats similaires à conditions de ne pas tronquer trop fort les données avant d'extraire des spectres. Il est préférable de tolérer la présence de contribution de sources dans les vecteurs lors de l'apprentissage de la structure de covariance que de tronquer trop fort les données au risque de biaiser l'estimation. La figure 3.20 regroupe toutes les estimations de la loi du maximum des spectres sous l'hypothèse $\mathcal{H}_0$. La courbe en pointillés (loi du maximum) représente ici la vraie

loi que nous cherchons à estimer à l'aide de l'apprentissage de la structure de covariance ou par le min-test. Nous observons que dans la zone d'intérêt, l'estimation par la covariance corrigée donne une courbe assez proche de la vérité. De même, **la distribution du minimum est une bonne approximation de la loi du maximum sous l'hypothèse nulle**. Cette méthode est entièrement non paramétrique, aucune hypothèse n'est faite sur la distribution des données autre que la symétrie du bruit avant filtrage, ce qui présente un avantage certain vis-à-vis des autres méthodes d'estimation. **Elle présente également l'avantage de prendre en compte la corrélation spatiale entre les spectres** (corrélation illustrée sur la figure 3.11b).

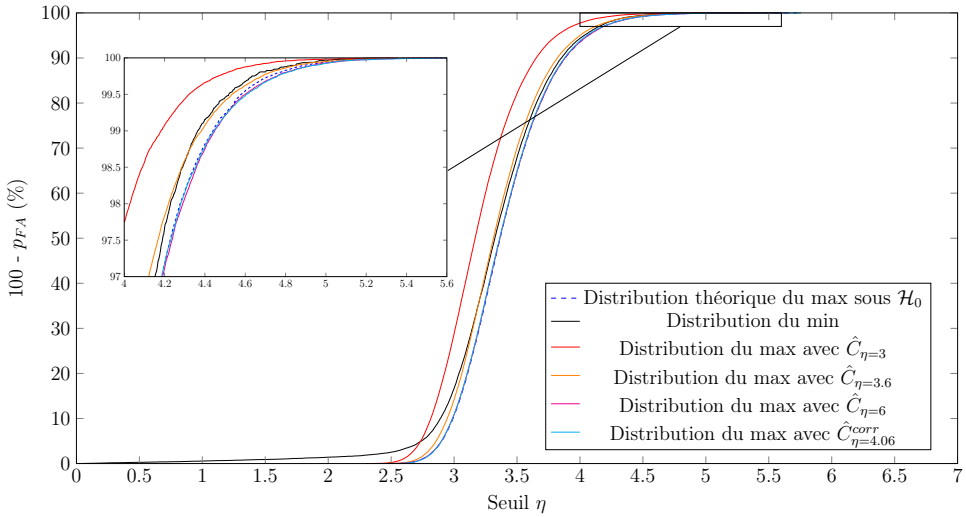


FIGURE 3.20 – Représentation des différentes estimations de la répartition de la loi du maximum des spectres avec différents niveaux de troncature des spectres pour estimer la loi du max-test obtenue sur les données i.i.d. corrélées par la matrice de covariance estimée (courbes rouge, orange et magenta). Ces courbes sont comparées à la loi théorique du maximum de données i.i.d. (courbe pointillée bleue) obtenue par méthode de Monte Carlo et à la distribution du minimum des spectres des données synthétiques (courbe noire).

3.3.4 Application du max-test aux données DryRun

Nous allons maintenant appliquer le max-test aux données DryRun. En appliquant le max-test à chaque spectre du cube, nous obtenons la carte des maxima qui est constituée de la valeur maximale de chaque spectre après filtrage adapté.

3.3.4.1 Loi des valeurs maximales des spectres

Afin de seuiller cette carte des maxima pour construire la carte de proposition, nous avons besoin de connaître la loi du max-test sous l'hypothèse $\mathcal{H}_0$ avec les méthodes décrites précédemment. Les courbes obtenues par la loi des minima et par apprentissage de la matrice de covariance après centrage des vecteurs sont comparées aux courbes des maxima et des minima sous l'hypothèse de bruit i.i.d. sur la figure 3.21. Notons que ces méthodes donnent des résultats équivalents à ceux obtenus sous l'hypothèse de bruit i.i.d. Cette hypothèse est en effet vérifiée sur les données DryRun. Cependant cette hypothèse est peu réaliste dans le cas des données réelles où l'apprentissage de $\mathcal{H}_0$ (paramétrique ou non) s'avère indispensable. Nous observons sur la figure 3.21 que la courbe de la répartition des valeurs minimales est légèrement biaisée, mais cela peut provenir du fait que parmi les 100×100 valeurs minimales utilisées pour estimer

cette courbe, un grand nombre sont contaminées par les composantes continues des galaxies dont le support spatial a été élargi par le filtrage adapté.

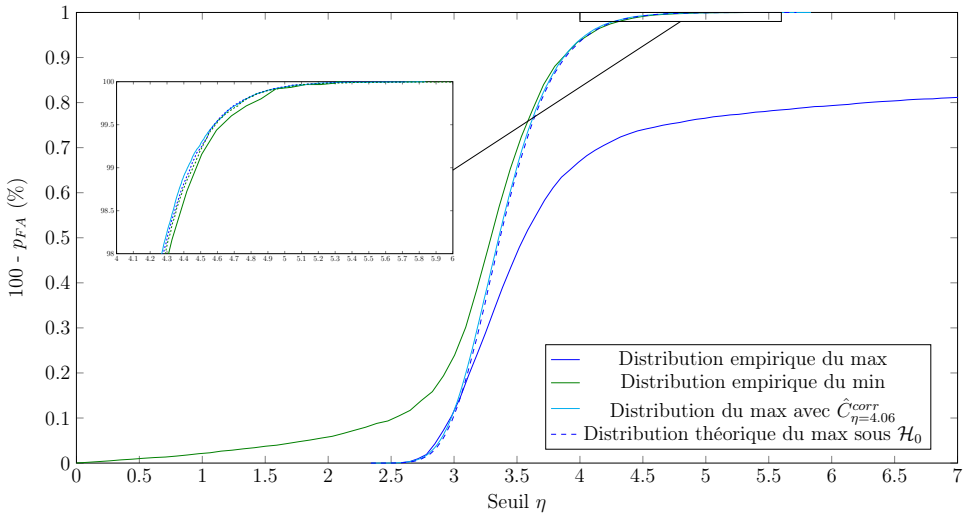


FIGURE 3.21 – Répartition empirique des valeurs maximales et des valeurs minimales calculées sur les spectres du DryRun après filtrage adapté. La répartition théorique du max-test sous l'hypothèse nulle est également reportée (courbe pointillée).

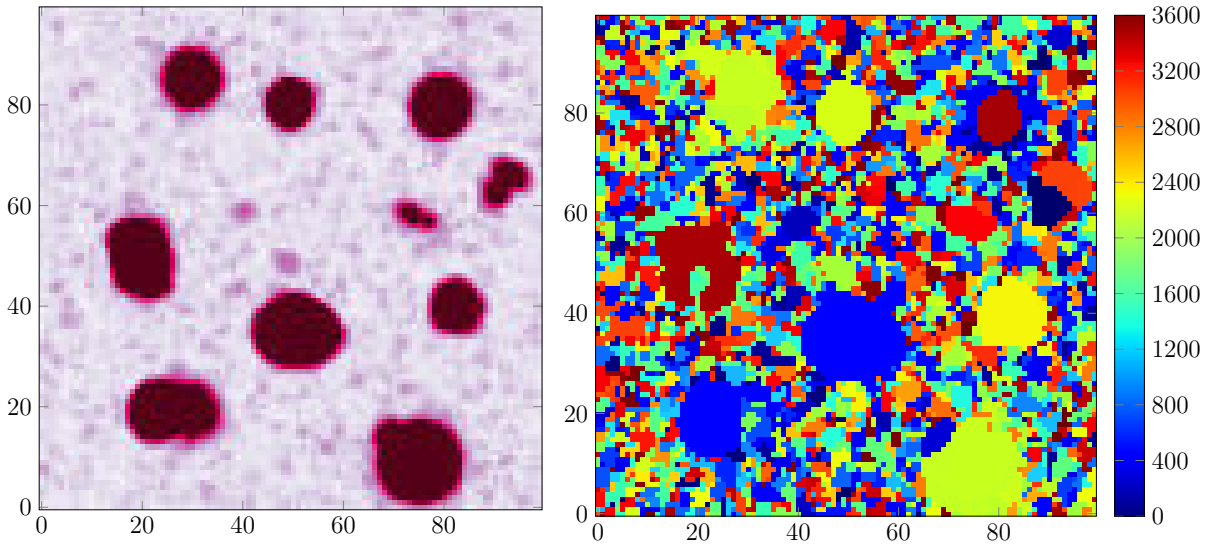
3.3.4.2 Carte des maxima et carte des longueurs d'onde

La carte des maxima obtenue est présentée sur la figure 3.22a, une carte complémentaire peut être produite à l'aide du max-test : la carte des longueurs d'onde qui indique la position dans le spectre de la valeur maximale de chaque spectre représentée sur la figure 3.22b. Mis à part pour les sources dont le spectre est constitué d'une composante continue qui oscille (spectres des étoiles par exemple, voir figure 1.8), nous pouvons constater l'homogénéité spectrale des galaxies ; en effet en comparant le support spatial des galaxies, figure 3.22a, avec les pixels correspondants sur la figure 3.22b, l'indice de la valeur maximale du spectre est constant dans le support de chaque source. Cette carte des indices peut être particulièrement utile pour séparer deux sources qui se recouvrent spatialement.

La carte des maxima peut maintenant être seuillée en fixant une probabilité de fausse alarme p_{FA} et en utilisant le seuil η correspondant donné par la courbe des valeurs minimale ou la courbe produite à l'aide de la matrice de covariance $\hat{C}^{corr}_{\eta=4.06}$. Par exemple pour une probabilité de fausse alarme $p_{FA} = 0.1\%$ le seuil vaut $\eta = 4.93$ et pour une probabilité de fausse alarme $p_{FA} = 1\%$ le seuil vaut $\eta = 4.45$.

3.3.4.3 Construction de la carte de proposition

A partir de la carte des maxima seuillée, nous obtenons une segmentation en deux classes, $\mathcal{C}_0$: les pixels de bruits au sens du max-test, $\mathcal{C}_1$: les pixels appartenant à une source. Il n'est pas envisageable de fournir tous les pixels classés dans $\mathcal{C}_1$. En effet, si l'algorithme accepte la naissance d'un objet en périphérie d'une source, l'énergie d'attache aux données est suffisamment forte pour empêcher la mort de l'objet mal placé. Nous pouvons envisager, en revanche, que l'objet soit translaté ; cependant, si d'autres objets ont été proposés entre temps en périphérie cette même galaxie, ils risquent de s'organiser en *grappe* et de donner lieu à des détections multiples. Nous avons vu au chapitre 2 que d'un point de vue interprétation physique, les détections multiples



(a) Carte des maxima.

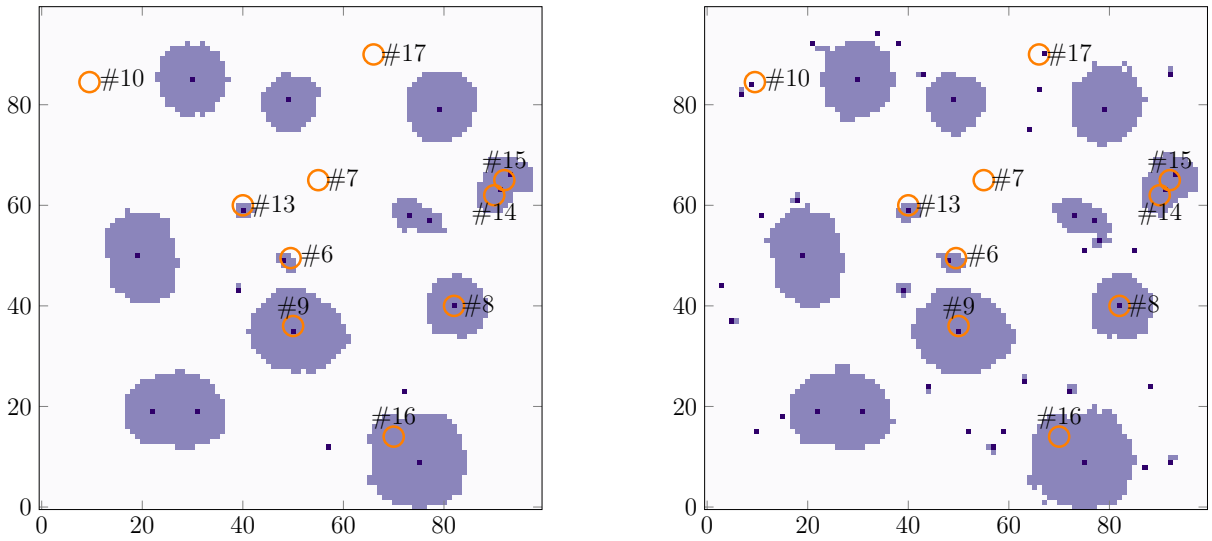
(b) Carte des longueurs d'onde

FIGURE 3.22 – Résultat du max-test, avec la carte des valeurs maximales des spectres après filtrage adapté (a) et la carte des longueurs d'onde qui indique la position dans le spectre de la valeur maximale de chaque spectre (b). Sur la carte des longueurs d'onde les positions sont données en indice des longueurs d'onde.

sont à éviter. **Pour éviter ce genre de phénomène, nous avons fait le choix de ne proposer que les maxima locaux des ensembles de pixels de la classe $\mathcal{C}_1$.**

Avec les deux valeurs de fausse alarme, $p_{FA} = 0.1\%$ et $p_{FA} = 1\%$, nous obtenons les cartes de proposition sur les figures 3.23a ($p_{FA} = 0.1\%$) et 3.23b ($p_{FA} = 1\%$). Les pixels blancs ont été classés dans $\mathcal{C}_0$, ils ne seront donc pas proposés comme centre de galaxies. Les autres pixels, de valeurs supérieures au seuil considéré, sont classés dans $\mathcal{C}_1$ (pixels violet clair et violet foncé sur les cartes 3.23a et 3.23b). Les pixels foncés sont les maxima locaux (au sens du 4-voisinage) sur la carte des valeurs maximales des spectres des ensembles de pixels de la classe $\mathcal{C}_1$. Ce sont ces pixels qui seront proposés comme centres de galaxies lors des mouvements de naissance de l'algorithme de détection proposé dans le chapitre 2. Les mouvements de translation sont autorisés tant que les centres restent localisés sur des pixels de la classe $\mathcal{C}_1$, en revanche un centre de galaxie ne peut être déplacé dans un pixel de la classe $\mathcal{C}_0$. C'est cette approche qui a été implémentée dans la première version de l'algorithme transféré au consortium MUSE.

Cependant avec une probabilité de fausse alarme 0.1%, 4 objets ne pourront être proposés : les objets #7, #10 et #17 qui sont les trois sources de plus faible RSB ($\leqslant -5dB$) et la galaxie #16 qui est située à proximité d'une étoile très brillante (#2). La source #16 n'est pas suffisamment brillante pour présenter un maximum local sur la carte des maxima. En revanche sur la carte des longueurs d'onde (figure 3.22b) les sources #2 et #16 sont clairement séparées. Afin de proposer tous les centres possibles (qui apparaissent sur la carte seuillée) il est possible d'envisager une stratégie différente : au lieu de ne proposer que les maxima locaux de la classe $\mathcal{C}_1$, il est possible d'exploiter les informations de seuillage et d'indice des longueurs d'onde. Pour cela, nous proposons de segmenter la carte seuillée à l'aide de la carte des indices de longueurs d'onde et de proposer pour les centres potentiels de galaxies le centroïde de chaque région segmentée. Ainsi l'ensemble de pixels formé par les sources #2 et #16 sera divisé en deux ensembles distincts. L'exploitation de la carte des longueurs d'onde fait partie des perspectives à développer pour affiner la carte de proposition. Il est aussi possible d'envisager de recalculer une nouvelle carte



(a) Carte de proposition $p_{FA} = 0.1\%$

(b) Carte de proposition $p_{FA} = 1\%$

FIGURE 3.23 – Cartes de proposition pour le cube DryRun pour des probabilités de fausse alarme de 0.1% (a) et 1% (b). Les pixels blancs ont été classés dans $\mathcal{C}_0$ des pixels de bruit, ils ne seront donc pas proposé comme centre de galaxies. Tous les autres pixels ont été classés dans $\mathcal{C}_1$ (pixels colorés), mais en tenant compte du fait que les galaxies présentent un maximum d'intensité en leur centre, nous sélectionnons tous les maxima locaux des ensembles de pixels de la classe $\mathcal{C}_1$ (pixels violet foncé) et ce sont ces pixels qui seront proposés comme centre de galaxies. Les centres sont ensuite autorisés à translater dans les zones violet clair.

de proposition au cours de la détection en tenant compte des sources déjà détectées et peut-être améliorer la détectabilité de galaxies de faible intensité situées dans l'étendue spatiale de sources plus brillantes comme c'est le cas des sources #2 et #16.

3.4 Contrôle du FDR : seuillage de la sortie du filtrage adapté par contrôle du taux de fausses découvertes

Le test proposé sur les spectres du cube après filtrage adapté permet de contrôler le FWER pour chaque spectre, seulement cela ne permet pas de contrôler un taux d'erreur global dans la liste de pixels retournés à l'algorithme pour la proposition de centre de galaxie. Il peut être plus intéressant de garantir un taux de fausses découvertes dans cette liste, et donc de contrôler le FDR sur tout le cube.

3.4.1 Notations et formulation du problème

Le modèle d'observation décrit par l'équation (2.8) peut se réécrire comme la somme des contributions des sources observées $S_{obs,j}$ (voir équation (2.3)) et un bruit additif gaussien ϵ :

$$\mathbf{Y}(p,q,\lambda) = \sum_j S_{obs,j}(p,q,\lambda) + \epsilon_{Bg}(p,q,\lambda), \tag{3.15}$$

où $\epsilon_{Bg}(p,q,\lambda) \sim \mathcal{N}(0,1)$. Rappelons que $S_{obs,j}$ est la composition de la source S_j avec la réponse de l'instrument. Nous pouvons reformuler le modèle décrit par l'équation (3.15) sous forme

vectorisée avec les notations introduites dans l'annexe E :

$$\mathbf{Y}^{(v)} = \mathbf{H}\boldsymbol{a} + \epsilon_{Bg}^{(v)}, \tag{3.16}$$

avec $\boldsymbol{a} = [a_1, \cdots, a_N]^T$ le vecteur d'intensité où chaque élément a_i est obtenu en sommant les contributions des différentes sources $S_j^{(v)}$ à la position (vectorisée) $i = 1, \cdots, N$, avec $N = P \times Q \times \Lambda$:

$$a_i = \sum_j S_j^{(v)}(i).$$

Le terme $\mathbf{H}\boldsymbol{a}$ de l'équation (3.16) résume ainsi la convolution des sources avec la PSF de l'instrument. Pour la suite de la méthode, nous avons besoin de définir deux hypothèses :

H1. La contribution des sources est positive, *i.e.* pour tout $i = 1, \cdots, N$, et pour tout $j = 1, \cdots, p : S_j^{(v)}(i) > 0$.

H2. La PSF est non négative, *i.e.* la matrice $\mathbf{H}$ est non négative, *i.e.* toutes les composantes de la matrice sont non négatives : $\mathbf{H}_{i,j} \geqslant 0$.

Dans le cas des données MUSE, ces deux hypothèses sont vérifiées.

Le nombre de sources et leur position sont *a priori* inconnus ; il faut donc tester les N positions possibles dans le champ de données. Tester la présence de sources à la position i revient à tester les N valeurs a_i avec les hypothèses définies par :

$$\begin{cases} \mathcal{H}_0^i & : & a_i = 0 & \text{(bruit seul)} \\ \mathcal{H}_1^i & : & a_i > 0 & \text{(source + bruit)} \end{cases} \tag{3.17}$$

Afin d'améliorer la détectabilité des sources, nous effectuons un filtrage adapté à la PSF de MUSE qui s'écrit sous forme vectorisée :

$$\mathbf{H}^T\mathbf{Y} = \mathbf{H}^T\mathbf{H}\boldsymbol{a} + \mathbf{H}^T\epsilon \tag{3.18}$$

La matrice $\mathbf{H}$ présente des propriétés importantes pour la suite du processus de détection :

P1. $\mathbf{H}$ est creuse, *i.e* elle contient peu de coefficients significatifs.

P2. $(\mathbf{H}^T\mathbf{H})_{i,i} = 1 \; \forall \; 1 \leqslant i \leqslant N$.

P3. $\mathbf{H} \geqslant 0$ est une matrice non négative.

D'après l'équation (3.18) et le modèle de mélange linéaire gaussien défini par l'équation (3.16), le vecteur $\mathbf{Y}^{f,(v)} = \mathbf{H}^T\mathbf{Y}$ est gaussien :

$$\mathbf{Y}^{f,(v)} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{3.19}$$

avec $\boldsymbol{\mu} = \mathbf{H}^T\mathbf{H}\boldsymbol{a} \geqslant 0$ et $\boldsymbol{\Sigma} = \mathbf{H}^T\mathbf{H} \geqslant 0$.

3.4.2 Formulation du test appliqué à chaque pixel

Le filtrage adapté modifie la formulation des tests décrits par l'éq. (3.17), un plus grand nombre d'échantillons vont être classés dans $\mathcal{H}_1$ car le filtrage adapté élargit les sources. D'après l'équation (3.19), chaque composante $\mathbf{Y}^{f,(v)}(i)$ de $\mathbf{Y}^{f,(v)}$ est gaussienne : $\mathbf{Y}^{f,(v)}(i) \sim \mathcal{N}(\mu_i, 1)$, avec μ_i la i^{eme} composante du vecteur $\boldsymbol{\mu}$, pour tout $1 \leqslant i \leqslant N$. La détection des contributions des sources dans la nouvelle statistique $\mathbf{Y}^{f,(v)}$ conduit alors au test suivant

$$\begin{cases} \mathcal{H}_0^i & : & \mu_i = 0 & \text{(bruit seul)} \\ \mathcal{H}_1^i & : & \mu_i > 0 & \text{(contribution d'une source)}, \end{cases} \tag{3.20}$$

3.4.2.1 P-valeurs et procédure BH

Dans le cas de n tests indépendants, la procédure BH proposée dans Benjamini and Hochberg [1995] contrôle le FDR à un niveau $\pi_0 q$ où $\pi_0 = \frac{N_0}{N}$ et N_0 est le nombre de tests réellement sous $\mathcal{H}_0$ et $0 \leq q \leq 1$ est le paramètre de contrôle. L'encadré 3.7 rappelle brièvement la procédure BH.

ENCADRÉ 3.7 – Procédure de contrôle du FDR de Benjamini-Hochberg

1. Soient $p_{(0)} < p_{(1)} \leqslant p_{(2)} \leqslant \cdots \leqslant p_{(N)}$ les N p-valeurs ordonnées, avec par convention $p_{(0)} = 0$.
2. Soit $k = \underset{i}{\text{argmax}} \left(p_{(i)} \leqslant q\frac{i}{N}\right)$
3. Rejet des hypothèses $\mathcal{H}_0^{(1)}, \cdots, \mathcal{H}_0^{(k)}$.

Dans le problème défini par l'éq. (3.20), les p-valeurs sont calculées à partir de la fonction de répartition $F_{\mathcal{H}_0}$ de la loi de $\mathbf{Y}^{f,(v)}$ sous $\mathcal{H}_0$ (ici $F_{\mathcal{H}_0} = \Phi$, la fonction de répartition de la loi normale) : $p_i = 1 - F_{\mathcal{H}_0}(\mathbf{Y}_i^{f,(v)})$. Par construction les p-valeurs p_i sont uniformément distribuées sur $[0, 1]$ sous $\mathcal{H}_0$. La fonction de répartition $F_{\mathcal{H}_0}$ est stochastiquement plus grande que celle sous $\mathcal{H}_1$, $F_{\mathcal{H}_0}(x) \geqslant F_{\mathcal{H}_1}(x)$.

Dans le cas du filtrage adapté, les $\mathbf{Y}^{f,(v)}(i)$ ne sont pas indépendants. Néanmoins sous certaines conditions de dépendance, pour utiliser directement la procédure BH pour seuiller les données, le $\mathbf{Y}^{f,(v)}$ doit vérifier la condition PRDS énoncée par Benjamini and Yekutieli [2001]. Cette propriété PRDS, pour *positive regression dependency on each one from a subset* I_0 (abrégé PRDS sur I_0), est définie ainsi :

Propriété PRDS : *Pour tout ensemble croissant D, et pour chaque $i \in I_0$, la probabilité $(Pr(X \in D | X_i = x)$ est non décroissante en x.*

L'ensemble D est dit croissant si $x \in D$ et $y \geqslant x$ implique $y \in D$. Le théorème principal de l'article est le suivant :

Théorème : *Si la distribution de la statistique de test est PRDS sur le sous-ensemble de tests correspondant à une vraie hypothèse nulle, alors la procédure de Benjamini and Hochberg [1995] contrôle le FDR à un niveau inférieur ou égal à $\frac{N_0}{N} q$.*

Dans l'exemple du cas gaussien multivarié proposé par Benjamini and Yekutieli [2001], les données sont PRDS si les éléments de la matrice de covariance, sous l'hypothèse $\mathcal{H}_0$, sont positifs.

3.4.2.2 Filtrage adapté et contrôle du FDR

D'après les travaux de Benjamini and Yekutieli [2001], si la distribution des p-valeurs est PRDS alors la procédure BH contrôle le FDR à un niveau inférieur ou égal à $\pi_0 q$. De plus, lorsque $\mathbf{\Sigma} \geqslant 0$ alors $\mathbf{Y}^{f,(v)} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$ est PRDS, et si f est une fonction monotone alors le vecteur $\mathbf{Y}^{f,(v)} = (f(\mathbf{Y}_1^{f,(v)}), \cdots, f(\mathbf{Y}_N^{f,(v)}))$ est aussi PRDS. Nous proposons de retranscrire les résultats de Benjamini and Yekutieli [2001] pour un problème emblématique du signal, la détection de sources par filtrage adapté, à l'aide de la procédure de seuillage par contrôle du FDR :

Proposition. *La procédure de Benjamini-Hochberg permet de seuiller la sortie du filtrage adapté tout en contrôlant le taux de fausses découvertes (FDR).*

Dans le cas qui nous intéresse, le résultat du filtrage adapté $\mathbf{Y}^{f,(v)} = H^T Y$ est en effet PRDS puisqu'il est gaussien avec une matrice de covariance non-négative, voir équation (3.19). Les p-valeurs se déduisant par une fonction décroissante du vecteur gaussien $\mathbf{Y}^{f,(v)}$ sont donc également PRDS.

La procédure BH permet de contrôler le FDR au niveau $q\pi_0 \leqslant q$ en seuillant les p-valeurs correspondantes. Notons que si π_0 n'est pas connue, le contrôle se fait au niveau q. Finalement, rejeter $\mathcal{H}_0^{(1)}, \cdots, \mathcal{H}_0^{(k)}$ est équivalent à dire que les composantes correspondantes, $\mathbf{Y}_{(1)}^{f,(v)}, \mathbf{Y}_{(2)}^{f,(v)}, \cdots, \mathbf{Y}_{(k)}^{f,(v)}$, sont significatives, *i.e* $\mu_{(1)} > 0, \mu_{(2)} > 0, \cdots, \mu_{(k)} > 0$.

3.4.3 Application au cube DryRun

Nous allons maintenant appliquer la méthode de seuillage par contrôle du FDR sur les données synthétiques DryRun. Le résultat du seuillage est un cube binaire, ce type de résultat est intéressant, notamment dans le cas où deux sources parfaitement superposées[6] dont le spectre contiendrait principalement une raie d'émission, mais à des longueurs d'onde différentes seraient détectables de façon séparée. Seulement, pour des raisons purement mathématiques, la méthode de détection proposée dans le chapitre 2 n'autorise pas la modélisation de deux sources superposées par deux objets distincts. Nous travaillerons donc sur une projection du cube binaire, afin d'obtenir une carte de proposition en deux dimensions.

3.4.3.1 Choix de la normalisation des données

Lorsque le cube de variance Σ_{MUSE} fourni avec le cube de données MUSE n'est pas fiable, la normalisation des données se fait uniquement à l'aide des moyennes et variances estimées par σ-clipping par point fixe. Dans ce cas, la sortie du filtrage adapté est un vecteur gaussien multivarié de matrice de covariance positive, la condition PRDS est respectée, et la procédure de BH peut être utilisée pour contrôler le FDR. Ce sera le cas du cube DryRun.

Si le cube de variance est fiable, il est utilisé pour réduire les données avant les opérations de réduction par σ-clipping par point fixe. Le cube de variance est un estimateur de la variance en chaque point du cube obtenu à partir des N_{poses} individuelles utilisées pour former le cube. Dans ce cas, le vecteur $\tilde{\mathbf{Y}}^{(v)} = \mathbf{Y}^{(v)}/S$, où $S = \Sigma_{MUSE}^{\frac{1}{2}}$ est la racine carrée du cube de variance Σ_{MUSE} vectorisé (cube d'écart-type vectorisé), est une version studentisée à N_{poses} degrés de liberté du vecteur gaussien multivarié $\mathbf{Y}^{(v)}$. Nous discutons ce cas dans le paragraphe 4.3.

3.4.3.2 Seuillage

Dans ce paragraphe nous allons appliquer la méthode de seuillage du cube de données DryRun à l'aide de la procédure BH décrite dans l'encadré 3.7. Lors de l'étape de seuillage du cube, si un pixel est classé sous $\mathcal{H}_1$ alors sa valeur est mise à 1, s'il est classé sous $\mathcal{H}_0$ alors sa valeur sera mise à 0. Nous pouvons ainsi espérer que dans le cas d'un spectre appartenant à une galaxie avec une composante spectrale continue, la somme des éléments de ce spectre sera proche de Λ, le nombre de bandes spectrales dans un cube MUSE. Dans le cas d'une galaxie de type Lyα, nous espérons observer un certain nombre de pixels consécutifs classé sous $\mathcal{H}_1$.

La figure 1.6 présente les résultats obtenus par seuillage du cube DryRun à l'aide de la procédure BH. La position et la forme des objets du DryRun sont connues, il est ainsi possible de calculer la proportion effective de fausses découvertes pour les différents seuillages. Pour un contrôle du FDR à $q = 5\%$, 83.3% des pixels des sources présentes dans le cube ont été détectés et la proportion de fausses découvertes s'élève à 2.16%. Le nombre de pixels classés sous $\mathcal{H}_1$ représente seulement 6.9% des pixels du cube. Pour un contrôle du FDR à $q = 10\%$, 85.3% des

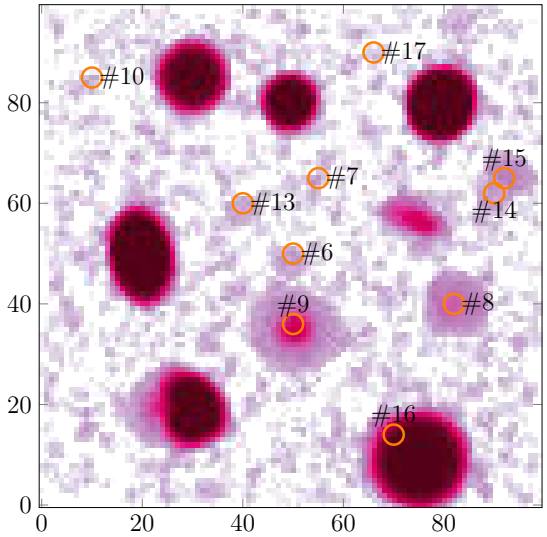
6. Leur centre est situé à l'intérieur du même pixel.

pixels des sources présentes dans le cube ont été détectés et la proportion de fausses découvertes s'élève à 4.42% Le nombre de pixels classés sous $\mathcal{H}_1$ représente seulement 7.3% des pixels du cube. Les galaxies de type Lyα sont toutes présentes (à l'exception de la galaxie #17) sur les cartes réduites (figures 3.24b et 3.24d). Bien que la proportion de fausses découvertes ne soit pas très élevée (inférieure à 1.4%) pour un contrôle du FDR inférieur à 10%, utiliser la carte obtenue en sommant le cube seuillé (figures 3.24a et 3.24c) ne réduira pas suffisamment la zone du cube à explorer. En effet, une grande majorité des positions spatiales (p, q) contiennent dans leur spectre au moins un pixel isolé classé sous $\mathcal{H}_1$. Nous pouvons raisonnablement espérer que ces spectres ne contiennent que du bruit.

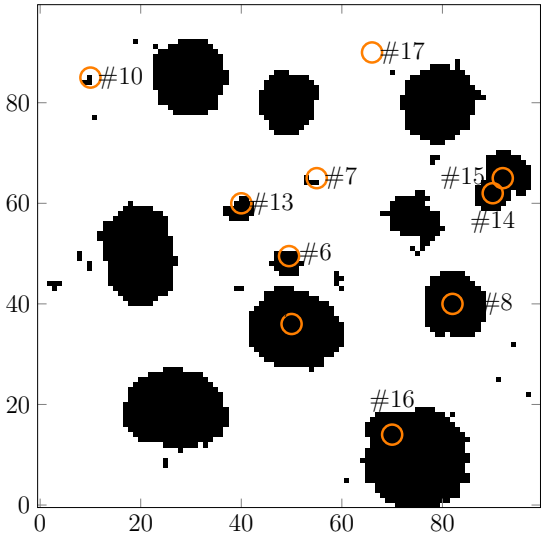
3.4.3.3 Carte de proposition

Une fois la carte binaire obtenue, il faut affiner la carte de proposition afin de donner à l'algorithme une liste de pixels candidats pour être des centres de galaxies. Là encore le problème se pose de savoir comment choisir ces pixels, il n'est pas envisageable de fournir tous les pixels classés dans $\mathcal{C}_1$. En retirant toutes les positions spatiales qui ne présentaient pas plus de 5 éléments consécutifs de leur spectre classés dans $\mathcal{C}_1$, nous restreignons une première fois la carte de proposition. Pour les raisons évoquées dans le paragraphe 3.3.4.3 il faut guider la proposition des centres de façon plus précise qu'en se restreignant aux cartes (figures 3.24b et 3.24d).

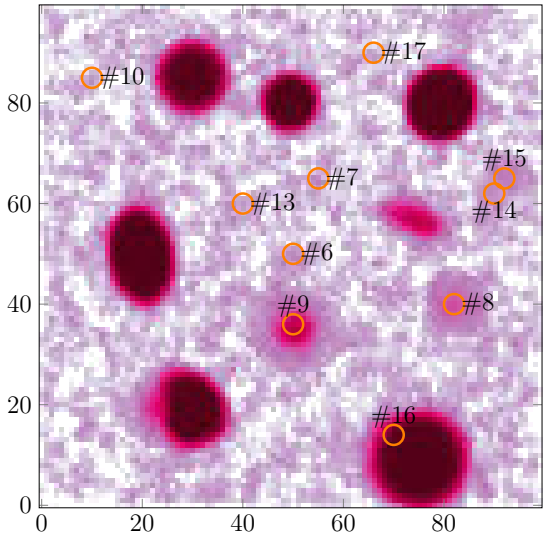
L'idée la plus simple est de proposer le centroïde de chaque ensemble de pixels de la classe $\mathcal{C}_1$. Seulement, le problème soulevé dans le paragraphe 3.3.4.3 pour les sources #2 et #16 se poserait à nouveau, de même pour les objets #14 et #15. Une seconde idée serait d'exploiter une carte de longueurs d'onde comme pour le max-test pour affiner la proposition des centres de galaxies. Cette carte de longueurs d'onde peut être construite en prenant l'indice de la longueur d'onde correspondant à la plus petite p-valeur de chaque spectre. Cette carte est représentée sur la figure 3.25. en utilisant les informations de longueurs d'onde, il serait possible de séparer les #2 et #16 et les sources #14 et #15.



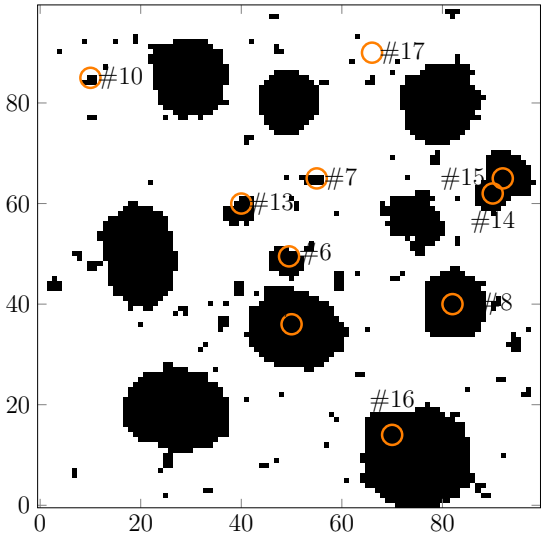
(a) Projection du cube seuillé (FDR = 5%).



(b) Réduction de la carte binaire (FDR = 5%)



(c) Projection du cube seuillé (FDR = 10%).



(d) Réduction de la carte binaire (FDR = 10%)

FIGURE 3.24 – Résultats du seuillage du cube DryRun par la procédure de contrôle du FDR par la procédure de Benjamini and Hochberg [1995] dans le cas de test dépendant. Les figures (a) et (c) sont les cartes obtenues en sommant les pixels (binaires) de chaque spectre. Les pixels blancs signifient que tous les pixels du spectre ont été classés dans $\mathcal{C}_0$, plus la couleur est foncée, plus le spectre contient de pixels classés sous $\mathcal{H}_1$. Les figures (b) et (d) tiennent compte des a priori physique que nous avons sur les galaxies de type Lyα : une raie d'émission a une largeur significative qui est étendue par la LSF lors du filtrage adapté. Tous les spectres qui ne présentent pas plus de 5 éléments consécutifs classés sous $\mathcal{H}_1$ sont rejetés de la carte de proposition. Les cercles oranges modélisent les galaxies de type Lyα dans les données DryRun.

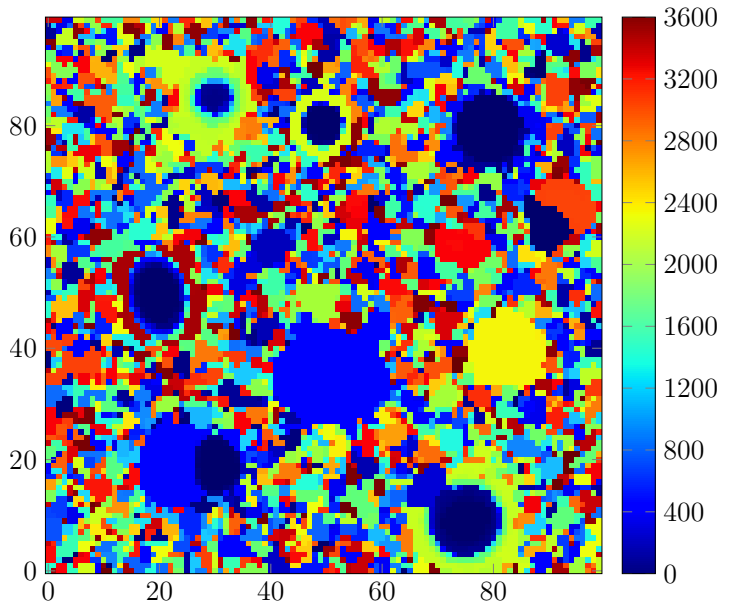


Figure 3.25 – Carte des longueurs d'onde correspondant à la position de la plus petite p-valeur de chaque spectre. L'échelle de couleurs est donnée en terme d'indice des longueurs d'onde.

3.5 Bilan

Filtrage adapté

Le filtrage adapté à la PSF, spectralement élargie, est une étape nécessaire à la prédétection des galaxies lointaines, mais cela introduit des corrélations spatiales et spectrales des données, y compris sous l'hypothèse nulle : les données ne contiennent que du bruit.

Contrôle des erreurs

Il existe deux critères de contrôle des erreurs dans le cas de tests multiples :

- le FWER (max-test, HC) qui permet de contrôler la probabilité de faire au moins une erreur de type 1 (fausse alarme) parmi les N tests,
- le FDR (procédure de Benjamini-Hochberg) qui permet de contrôler la proportion de fausses découvertes parmi toutes les découvertes.

Carte de proposition

La carte de proposition a une double utilité :

- elle guide les mouvements de naissance de l'algorithme RJMCMC en favorisant les pixels les plus susceptibles d'être un centre de galaxie,
- elle permet de définir la fonction d'intensité du processus ponctuel en interdisant certaines zones du cube qui ont été classées sous $\mathcal{H}_0$ par le test qui a servi à construire la carte.

Chapitre 4

Application aux données réelles

Sommaire

Ce chapitre est exclusivement dédié à l'application des méthodes proposées dans les chapitres 2 et 3 au cube de données réelles HDFS décrit dans le paragraphe 1.4.2. Le but est de mener l'étude détaillée des différentes étapes du processus de détection, du prétraitement des données à la production d'un catalogue d'objets détectés.

Dans la première partie de ce chapitre, nous nous intéresserons à l'impact des prétraitements sur les données, en termes de normalisation des données à partir des estimateurs de moyenne

et de variance sur chaque image du cube, et d'amélioration du RSB des objets par le filtrage adapté.

Dans la deuxième et la troisième parties de ce chapitre, nous nous intéresserons à la construction de la carte de proposition par le max-test et par le contrôle du FDR. Les techniques d'estimation de la loi des valeurs maximales des spectres après filtrage adapté seront mises en oeuvre sur le cube HDFS afin de seuiller la carte des valeurs maximales après filtrage adapté du cube de données.

Nous verrons ensuite comment initialiser la détection des galaxies dont le spectre contient principalement une raie d'émission, en effectuant, au préalable, la détection des sources à spectre continu sur l'image blanche.

Enfin nous analyserons le résultat de la détection de galaxies sur le cube HDFS en comparant le catalogue d'objets détectés à celui des sources détectées sur l'image du HDFS observée par le télescope spatial Hubble.

4.1 Prétraitement du cube de données

Le cube de données HDFS est le résultat d'une combinaison de 54 poses individuelles, *i.e.* des cubes de données de mêmes dimensions que le cube final. Si ce processus permet de collecter la lumière de galaxies très peu brillantes, il nécessite, en contre-partie, un certain nombre de traitements pour aligner les différentes poses, réduire les décalages d'intensité entre les capteurs (variation de température entre les poses et entre les différents capteurs au sein d'une même pose, etc). Tous ces traitements ont des conséquences sur les données, et comme dans toute application réelle, nous allons voir que les hypothèses utilisées pour modéliser les données ne sont pas forcément vérifiées.

4.1.1 Le cube HDFS

Le cube utilisé dans ce manuscrit ne correspond pas à la version publique du cube[1]. Il s'agit de la version 1.24 qui présente moins de systématiques (structures dues à la réduction de données) que la version publique. Dans le cadre de la détection de galaxies avec la méthode proposée dans le chapitre 2, nous avons fait l'hypothèse que la variance du bruit est identique sur chaque plan monochromatique. Or en combinant plusieurs poses individuelles qui ne sont pas parfaitement alignées, la variance du bruit augmente naturellement sur les bords du cube final. Afin de respecter au maximum l'hypothèse de stationnarité du bruit, nous avons fait le choix de tronquer le cube HDFS. Le cube utilisé a une taille finale de $311 \times 311 \times 3641$ pixels, ce qui correspond à un champ d'observation de 62.2×62.2 arcsec^2 (le cube n'a pas été tronqué en longueurs d'onde).

4.1.2 Extraction d'images pour l'étude de l'influence des prétraitements

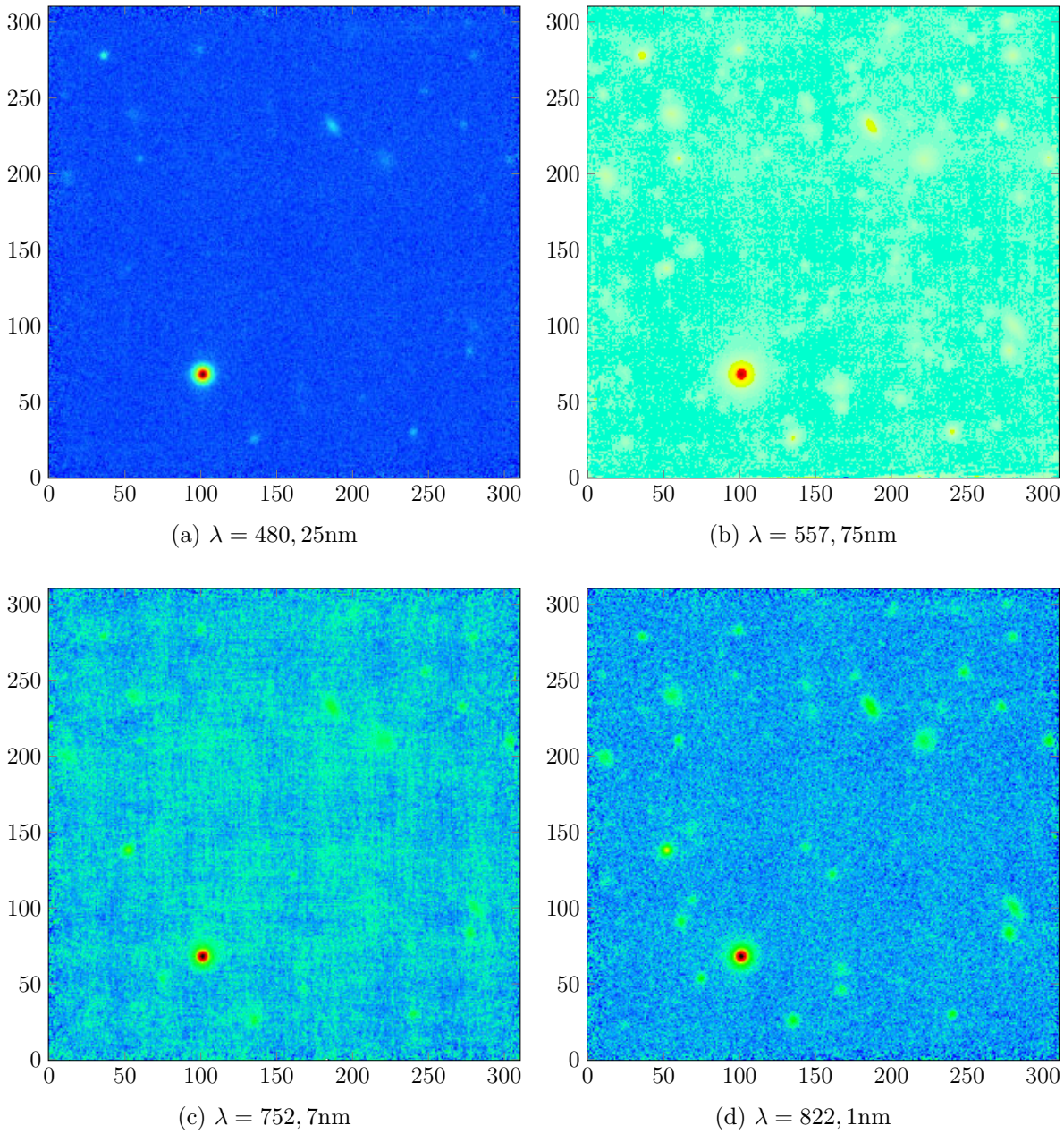
Afin d'étudier l'influence des différents prétraitements appliqués aux données, quatre images ont été extraites du cube de données HDFS. Ces images ont été sélectionnées à certaines longueurs d'onde d'intérêt :

- $\lambda = 480,25$nm : peu de galaxies, seulement celles dont le spectre contient une composante continue significative,
- $\lambda = 557,75$nm : longueur d'onde pour laquelle la variance des données est très grande,
- $\lambda = 752,7$nm : présence de fortes structures dans les données masquant une partie des galaxies,

1. Version 1.0 disponible à l'adresse : http://muse-vlt.eu/science/data-releases/

— $\lambda = 822,1$nm : présence de galaxies Lyman-alpha avec une raie d'émission étroite et significative à cette longueur d'onde.

Ces quatre images sont présentées sur la figure 4.1. Les sources qui apparaissent sur les quatre images sont des galaxies et des étoiles qui ont une composante continue significative dans leur spectre. Les sources qui n'apparaissent que sur une seule image sont des galaxies dont le spectre contient principalement une raie d'émission (Lyα ou autre).



(a) $\lambda = 480,25$nm

(b) $\lambda = 557,75$nm

(c) $\lambda = 752,7$nm

(d) $\lambda = 822,1$nm

FIGURE 4.1 – Images extraites du cube de données HDFS à différentes longueurs d'onde avant tout traitement. L'échelle de couleur varie car la dynamique des images est très différente d'une longueur d'onde à l'autre.

4.1.3 Estimation de la moyenne et de la variance

Sur la figure 4.1, toutes les images du cube ne sont pas affectées de la même façon par le bruit et les structures résiduelles de l'assemblage des données, issues des 24 spectrographes constituant MUSE. La présence de structures horizontales et verticales vont, bien entendu, à l'encontre des hypothèses de gaussiannité et de stationnarité spatiale du bruit. Nous devrons donc accepter de possibles erreurs de détection dues à la présence de ces structures et à l'impossibilité de centrer et réduire correctement les images localisées à certaines longueurs d'onde.

Sur la figure 4.2 sont présentés les histogrammes des intensités des quatre images introduites dans le paragraphe 4.1.2. On peut constater que pour les longueurs d'ondes $\lambda = 480.25$nm et $\lambda = 822.1$nm qui ne présentent pas de défaut apparent sur les images 4.1a et 4.1d, les trois estimateurs de moyenne et de variance donnent des résultats similaires qui permettent d'ajuster une loi gaussienne à l'histogramme des données (courbes noires en pointillés), sauf à droite de l'histogramme, où la contribution des sources alourdit la queue de la distribution. En revanche, pour les longueurs d'onde problématiques $\lambda = 557.75$nm (variance très grande) et $\lambda = 752.7$nm (présence de structures), les estimateurs de moyenne et de variance ne sont plus équivalents d'une méthode d'estimation à l'autre. Dans le cas où la variance des données est très grande (avant normalisation par le cube de variance Σ_{MUSE}), la méthode par σ-clipping implémentée dans *mpdaf* est biaisée, le mode de la courbe bleue est décalé vers la droite (figure 4.2b) ce qui signifie que la troncature des données n'est pas suffisamment conservative : un grand nombre de pixels appartenant à des sources ont été pris en compte dans l'estimation de la moyenne et de l'écart-type. Dans le cas où l'image est contaminée par des structures horizontales et verticales (figure 4.1c), c'est l'estimation paramétrique qui est biaisée (figure 4.2c), sans doute par les intensités positives au niveau des structures. Dans ces deux cas, l'estimation par σ-clipping par point fixe donne les résultats qui permettent d'ajuster correctement les valeurs négatives ou de faibles amplitudes. L'estimation obtenue semble alors très satisfaisante à un facteur d'échelle près lié à l'estimation de $\hat{\pi}_0$ (la connaissance précise de ce facteur n'est pas nécessaire pour la suite des traitements).

Afin de centrer et réduire le cube de données, nous choisissons la méthode de σ-clipping par point fixe qui donne des résultats satisfaisants en terme de précision et de robustesse sur les quatre images. Cette méthode fournissait également les meilleures performances en terme de biais et d'erreur quadratique moyenne sur les données synthétiques dans le paragraphe 3.1.3.4.

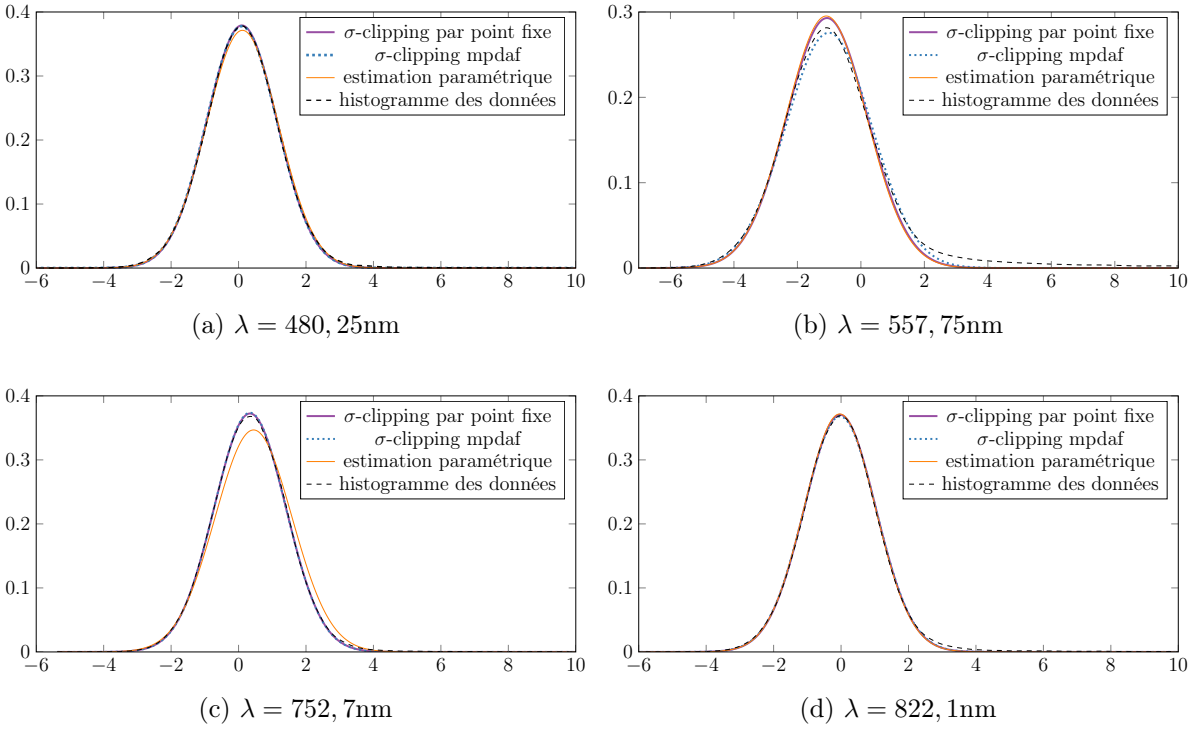
4.1.4 Filtrage adapté

Nous avons introduit dans le chapitre 3 la nécessité de réaliser un filtrage adapté à la PSF de l'instrument afin d'améliorer la détectabilité des galaxies lointaines, quasi-ponctuelles et de faible intensité.

Comme nous pouvons le constater en comparant les images avant filtrage, figure 4.1 et après filtrage, figure 4.3, le filtrage adapté étale la réponse spatiale (et spectrale) des galaxies, et augmente leur rapport signal à bruit. Tout signal dont la structure ressemble à la réponse de la PSF se retrouve ainsi amplifié. En revanche, le filtre adapté ne doit pas modifier la valeur moyenne (la combinaison linéaire de variables centrées reste centrée) et la variance (la PSF est ℓ_2-normalisée) des pixels de bruit si l'hypothèse de bruit gaussien centré-réduit i.i.d. est respectée.

Nous pouvons noter sur les quatre images présentées sur la figure 4.3 la présence de valeurs très proches de zéro (pixels bleu foncé) voire très négatives (pixels blancs). Ces ensembles de pixels ressemblent, au signe près, à la FSF de MUSE, et sont donc accentués par le filtrage adapté. Ceci entraîne la présence de valeurs négatives de forte amplitude qui décentrent les données après filtrage adapté.

De plus, nous observons dans les zones ne contenant pas de galaxies, que les données ne sont pas réduites, alors que sous l'hypothèse $\mathcal{H}_0$, pour un bruit gaussien centré-réduit i.i.d.,



(a) $\lambda = 480,25$nm (b) $\lambda = 557,75$nm

(c) $\lambda = 752,7$nm (d) $\lambda = 822,1$nm

FIGURE 4.2 – Histogramme des données (courbes pointillées noires) aux quatre longueurs d'onde considérées dans ce paragraphe (après normalisation par le cube de variance Σ_{MUSE}) et densités gaussiennes estimées $\hat{\pi}_0 f_0$.

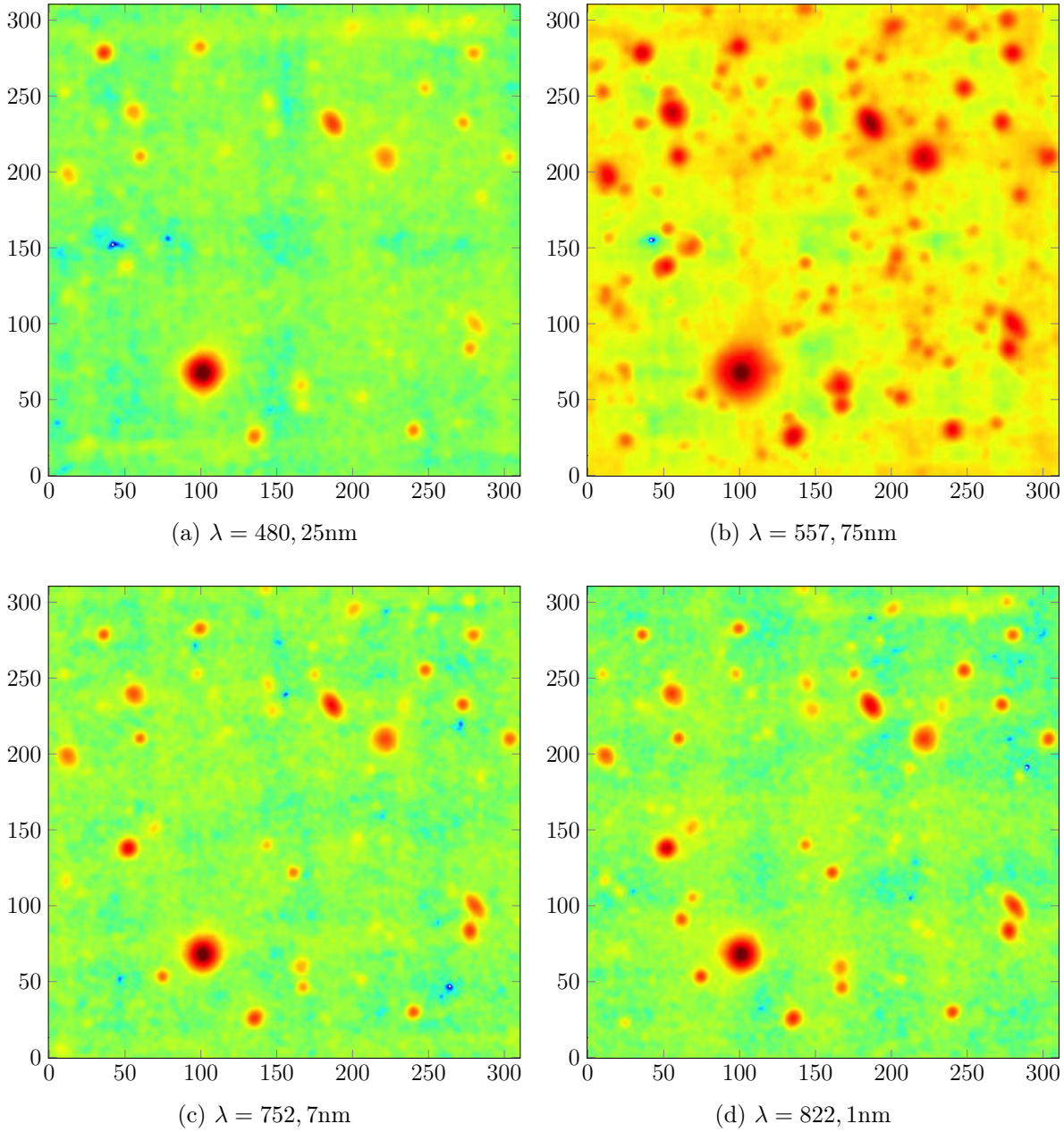
la sortie du filtrage adapté est un bruit gaussien centré-réduit. Nous savons que les données MUSE, avant filtrage adapté, sont localement corrélées par l'opération de drizzling. En effet, un cube de données réelles, tel que le HDFS, est issu de la combinaison de plusieurs dizaines de poses individuelles. Cette opération de drizzling entraîne une corrélation locale des pixels dans un voisinage en trois dimensions. Le plus petit voisinage 3D envisagé est un 26-voisinage [2], la corrélation peut s'étendre à un voisinage plus grand. Cette corrélation en trois dimensions des pixels voisins dans le cube final n'a pas d'expression analytique et elle est très dépendante des données, de la définition de la grille de pixels utilisée pour aligner les poses individuelles, etc. Tenir compte de ces corrélations en trois dimensions demanderait de propager pour chacun des 3×10^8 pixels du cube un sous-cube de covariance. Etant donnée la dimension des données MUSE, cette solution n'est pas envisageable. Nous verrons dans le paragraphe 4.2.1 comment essayer de tenir compte de cette corrélation.

Afin de toujours travailler sur des données centrées-réduites et se rapprocher du cas idéal où le bruit serait gaussien centré-réduit i.i.d., **nous normalisons la sortie du filtrage adapté avec les nouveaux estimateurs de moyenne et de variance fournis par le σ-clipping par point fixe**. La figure 4.4 illustre l'ensemble des traitements appliqués aux données dans la phase de prétraitement avant la construction de la carte de proposition.

Il faut noter que le filtrage adapté élargit la réponse des sources et réduit donc le nombre de pixels utilisables pour estimer la moyenne et la variance des pixels de bruit d'une image. Typiquement, 77% des pixels d'une image sont conservés pour réaliser l'estimation avant filtrage adapté contre 75% après. La différence n'est pas conséquente, même si la composante spatiale du filtrage adapté élargit les réponses des sources d'une dizaine de pixels [3] (rayon de la FSF).

2. Extension du 8-voisinage défini dans $\mathbb{R}^2$ à $\mathbb{R}^3$, on considère alors un cube de $3\times3\times3$ pixels centré sur le pixel d'intérêt.

3. La largeur à mi-hauteur de la FSF est d'environ 3 pixels, le profil Moffat décroit ensuite lentement vers 0.



(a) $\lambda = 480,25\text{nm}$ (b) $\lambda = 557,75\text{nm}$

(c) $\lambda = 752,7\text{nm}$ (d) $\lambda = 822,1\text{nm}$

Figure 4.3 – Images extraites du cube de données HDFS à différentes longueurs d'onde après filtrage adapté. L'échelle de couleur varie car la dynamique des images est différentes d'une longueur d'onde à l'autre du fait de la corrélation (spatiale et spectrale) du bruit, les données en sortie du filtrage ne sont pas centrées réduites.

Nous vérifions la validité de l'hypothèse de gaussiannité, sous l'hypothèse de bruit seul, des données du HDFS en traçant les diagrammes quantiles-quantiles des données à chaque étape du prétraitement sur la figure 4.5. Si les points bleus sont alignés sur la courbe $y = x$ alors la distribution des données suit probablement une loi normale. Nous observons un décrochage sur la droite du diagramme, en particulier pour les données après filtrage adapté, du fait de la présence de sources dans les données. Le léger décrochage sur la gauche du diagramme est dû à la présence de structures à valeurs négatives dans les données.

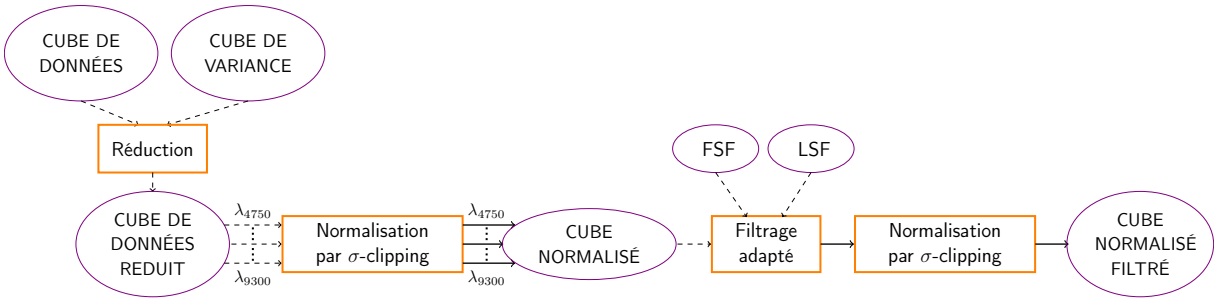
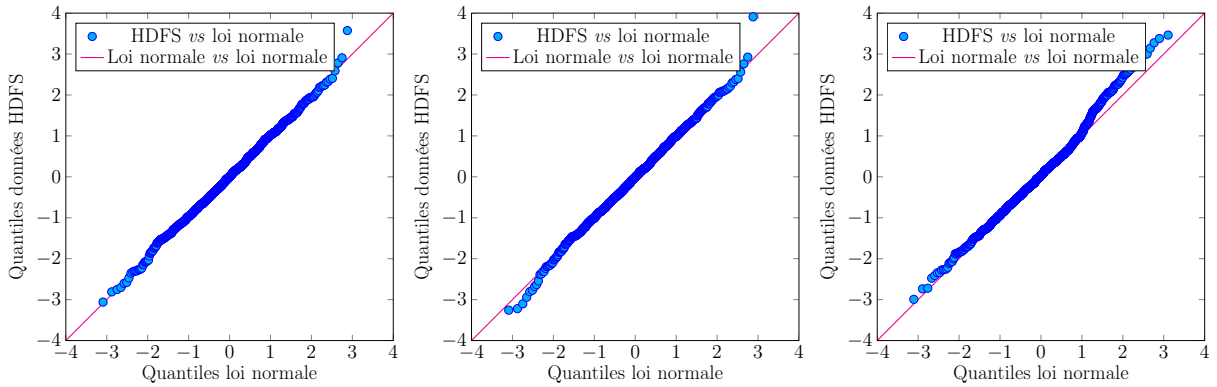


FIGURE 4.4 – Chaîne de prétraitement des données.



(a) (b) (c)

FIGURE 4.5 – Diagrammes quantiles-quantiles des données HDFS avant réduction par le cube de variance (a), après réduction par le cube de variance (b) et après filtrage adapté (c). A noter que dans les trois cas, les données sont centrées-réduites à l'aide des estimateurs de moyenne et de variance de la loi sous l'hypothèse $\mathcal{H}_0$ obtenus par σ-clipping par point fixe afin de comparer à la loi normale.

4.2 Construction de la carte de proposition par le max-test

Une fois les opérations de filtrage adapté, de centrage et de réduction effectuées sur les données réelles, nous pouvons appliquer le max-test défini dans le paragraphe 3.3. La loi du max-test donnée dans le paragraphe 3.3.3 est valable pour des données, sous l'hypothèse nulle, i.i.d. gaussiennes centrées et réduites. Nous savons que les données MUSE, même sous hypothèse de bruit seul, sont corrélées ; nous allons donc appliquer les différentes approches proposées dans le paragraphe 3.3.3 pour apprendre la loi du maximum des spectres dans le cas des données réelles. Le but de cette étude est de vérifier le comportement du max-test pour différents modèles de bruit pour les valeurs extrêmes du test sous $\mathcal{H}_0$.

4.2.1 Estimation de la loi du max-test

Afin d'utiliser un test comme le max-test présenté dans le paragraphe 3.3, il faut être en mesure de d'estimer la loi du test sous l'hypothèse nulle, *i.e.* lorsque le spectre considéré n'appartient pas à une source.

4.2.1.1 Modélisations des spectres sous l'hypothèse $\mathcal{H}_0$

Hypothèse de bruit gaussien i.i.d.

L'hypothèse la plus simple, celle qui nous a permis d'écrire le modèle d'observation pour la méthode de détection des galaxies présentée au chapitre 2, est celle de bruit i.i.d. gaussien centré réduit, *i.e.* sous l'hypothèse nulle, la distribution d'un spectre $\mathbf{y}_r$ à la position $r \equiv (p, q)$ peut s'écrire :

$$\mathbf{y}_r \sim \mathcal{N}(0, \mathbf{I}_\Lambda), \tag{4.1}$$

avec $\mathbf{I}_\Lambda$ la matrice identité de taille $\Lambda \times \Lambda$. Après filtrage adapté cette hypothèse peut se reformuler de la façon suivante :

$$\mathcal{H}_0^{norm.\ i.i.d} : \mathbf{y}_r^f \sim \mathcal{N}(0, C), \tag{4.2}$$

où C est la matrice de covariance des spectres après filtrage adapté dont l'expression est donnée par l'équation 3.12. Cette hypothèse simple n'est pas réaliste puisqu'elle ne prend pas en compte la corrélation des pixels de bruit avant le filtrage adapté.

Hypothèse de bruit gaussien corrélé

Afin de prendre en compte la corrélation des pixels sur les données réelles sous l'hypothèse nulle, nous choisissons d'estimer directement la structure de corrélation sur les données après filtrage adapté. Intuitivement, la corrélation des pixels dans un spectre de bruit ne devrait pas dépasser la taille de la LSF élargie (soit 19 pixels) et quelques pixels supplémentaires pour la corrélation introduite par l'étape de drizzling. Nous choisissons d'apprendre la structure de corrélation globale (qui prend en compte la corrélation des pixels de bruit dans un spectre sous $\mathcal{H}_0$ et la corrélation introduite par le filtrage adapté) à l'aide des méthodes présentées dans les paragraphes 3.3.3.3 et 3.3.3.4 sur des vecteurs de taille $d = 25$. Nous avons vu sur les données synthétiques que l'apprentissage de la corrélation en utilisant des vecteurs centrés donnent les meilleurs résultats en terme de biais d'estimation de la structure de corrélation. Cependant aucune de ces méthodes, y compris cette dernière, ne permet d'estimer correctement la structure de corrélation. Au lieu de retrouver une corrélation qui décroit vers zéro au fur et à mesure que l'on s'éloigne de la diagonale, les valeurs restent très élevées (> 0.5). Ceci est probablement dû à une composante continue qui reste présente même après centrage. N'ayant pas réussi à obtenir une meilleure estimation, nous utiliserons l'estimation de la matrice de covariance $\hat{C}_{\eta=4}^{corr}$ définie par l'équation (3.14) pour construire la loi empirique de la valeur maximale des spectres sous l'hypothèse de bruit gaussien corrélé.

Hypothèse de bruit de student i.i.d.

En réalité, le cube de données HDFS a été réduit par le cube de variance qui a été estimé à partir des $N_{poses} = 54$ poses individuelles qui ont servi à construire le cube. Les pixels du cube de variance sont distribués selon une loi du χ^2 à $N_{poses} - 1$ degrés de liberté. Les pixels du cube après réduction suivent donc une loi de Student à $N_{poses} - 1$ degrés de liberté. Notons que la loi de Student à $N_{poses} - 1 \simeq 50$ [4] degrés de liberté est proche de la loi gaussienne, les deux lois diffèrent légèrement dans les queues des distributions, voir la figure 4.6. Cette figure montre bien que la loi gaussienne est une très bonne approximation jusqu'au quantile 10^{-4} (ou 1- 10^{-4} par symétrie), et reste acceptable jusqu'à 10^{-5}.

4. Le nombre de poses individuelles prise en compte pour la construction du cube varie légèrement d'un pixel à l'autre, puisque certaines poses ont pu être écartées par le σ-clipping visant à éliminer les pixels contaminés par des rayons cosmiques.

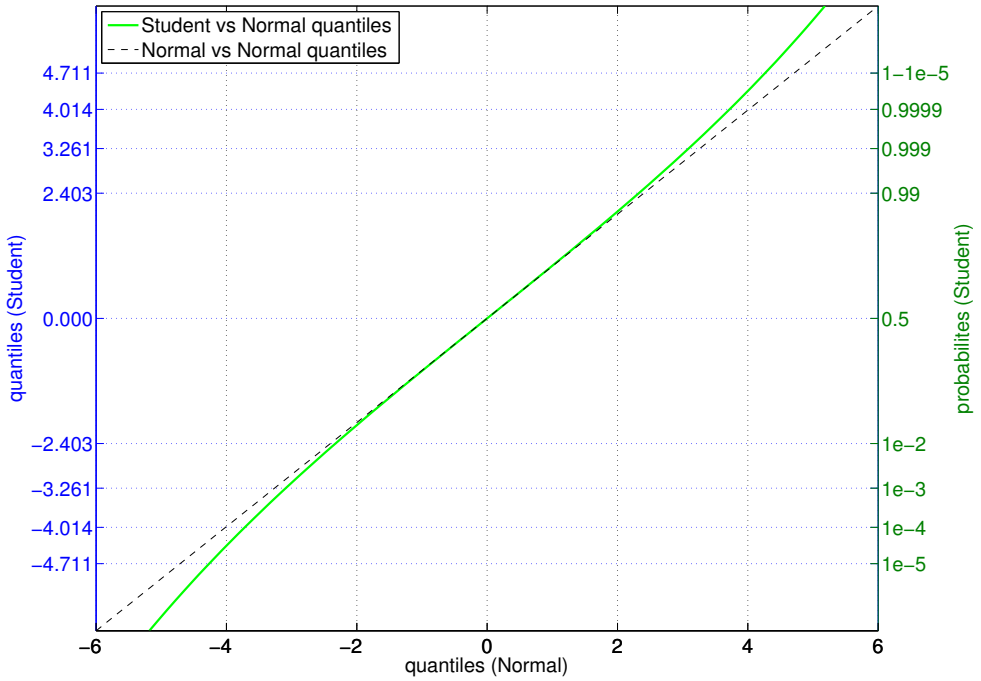


FIGURE 4.6 – Représentation par qq-plot de la pertinence de l'approximation de la loi de Student à 50 degrés de liberté par la loi gaussienne.

Hypothèse de bruit de student corrélé

De même que pour l'hypothèse de bruit gaussien corrélé, nous utiliserons la matrice $\hat{C}^{corr}_{\eta=4}$ pour construire la loi empirique de la valeur maximale des spectres sous l'hypothèse de bruit de student corrélé.

4.2.1.2 Comparaison des différentes lois obtenues pour la valeur maximale des spectres sous l'hypothèse nulle.

Nous représentons sur la figure 4.7 la loi empirique des valeurs minimales (en inversant le signe) des spectres du cube, ainsi que les lois du maximum des spectres obtenues pour les différentes hypothèses de distribution des pixels d'un spectre sous l'hypothèse nulle.

Comme sur les données synthétiques, le test basé sur les valeurs minimales des données (courbe noire sur la figure 4.7) est un peu moins conservatif que les autres tests, même si les lois s'avèrent similaires pour les petites probabilités de fausse alarme. Comme il est difficile de faire confiance aux autres tests (bruit normal ou de student, i.i.d. ou corrélés), nous faisons le choix d'utiliser la loi empirique des valeurs minimales pour calibrer le max-test. Cet estimateur non paramétrique de la loi de la statistique du max-test présente l'avantage de prendre en compte directement la structure de corrélation 3D sans devoir l'apprendre. Le tableau 4.1 présente la correspondance entre la probabilité de fausse alarme souhaitée pour le test et la valeur du seuil à utiliser pour construire ensuite la carte de proposition.

4.2.2 Carte de proposition

La carte de proposition se déduit de la carte des valeurs maximales des spectres seuillée à un seuil $\eta = 4.8$, correspondant à une probabilité de fausse alarme pour le max-test de $p_{FA} = 0.1\%$, d'après le tableau 4.1. Cette carte de proposition est présentée sur la figure 4.8

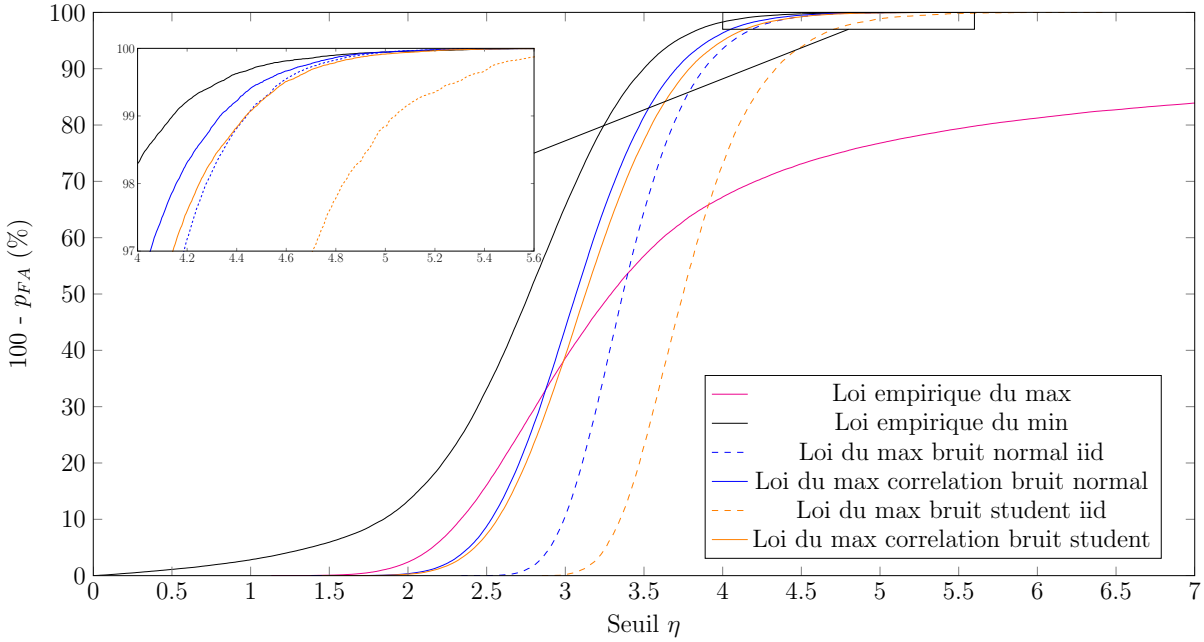


FIGURE 4.7 – Représentation des fonctions de répartition empiriques de la valeur maximale (courbe magenta) et de l'opposé de la valeur minimale (courbe noire) des spectres du cube HDFS après opération de filtrage adapté et de centrage et réduction. En bleu (resp. en orange) sont représentées les courbes obtenues par méthode de Monte Carlo sur des spectres de bruit i.i.d. distribués selon une loi normale (resp. distribués selon une loi de student) (courbes en trait plein) et des spectres corrélés à l'aide de la matrice de covariance apprises sur les spectres du cube HDFS (courbes en trait pointillé).

Un certain nombre de maxima locaux sur les bords de la carte (en haut à droite par exemple) sont dus à des structures de bruit présentant une variance plus élevée que le reste du cube et qui n'ont pas suffisamment été réduites par la variance estimée sur le cube lors de la phase de prétraitement. Nous savons en effet que les bordures du cube présentent un rapport signal à bruit moins élevé qu'au centre du cube car il y a moins d'observations disponibles (toutes les poses individuelles ne se superposent pas parfaitement) ce qui entraîne une plus grande variance.

Cette carte de proposition contient 301 pixels candidats pour être des centres de galaxies. Certains de ces pixels ont déjà été utilisés pour être des centres de galaxies à spectre continu lors de la détection sur l'image blanche (voir paragraphe 4.4) ; s'ils sont proposés lors de l'échantillonnage de la configuration d'objets sur le cube complet, les mouvements de naissance correspondants seront systématiquement rejetés si la détection a déjà été réalisée sur l'image blanche.

Cette carte binaire sert également à définir la fonction d'intensité du processus ponctuels de référence dans le modèle Bayésien. Au niveau des pixels classés dans $\mathcal{C}_0$, la fonction d'intensité sera nulle, ce qui signifie qu'aucun point (centre de galaxie) ne peut être placé sur ces pixels.

4.2.3 Carte des longueurs d'onde

La carte des longueurs d'onde, présentée sur la figure 4.9, indique la position de la valeur maximale de chacun des spectres du cube après filtrage adapté.

p_{FA}	η
10 %	3.49
5 %	3.70
3 %	3.85
2 %	3.96
1 %	4.15
0.5 %	4.33
0.1 %	4.8

TABLEAU 4.1 – Correspondance entre la probabilité de fausse alarme et le seuil à utiliser pour le max-test

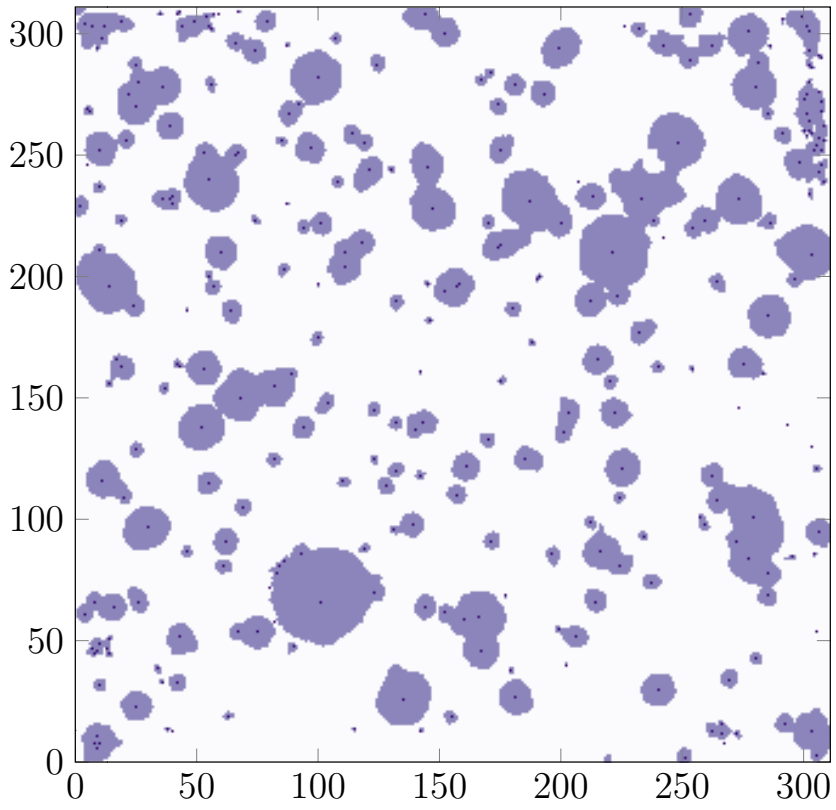


FIGURE 4.8 – Carte de proposition obtenue en appliquant le max-test à chaque spectre pour un seuil $\eta = 4.8$ correspondant à une probabilité de fausse alarme $p_{FA} = 0.1\%$ pour la loi des valeurs minimales du cube HDFS. Les pixels blancs ne seront jamais proposés, ils ont été classés dans $\mathcal{C}_0$ par le max-test. Les pixels colorés ont tous été classés dans $\mathcal{C}_1$ par le max-test avec une probabilité de fausses alarmes $p_{FA} = 0.1\%$. A noter que les pixels violet foncé sont les maxima locaux des ensembles de pixels classés dans $\mathcal{C}_1$. Seuls ces pixels seront proposés comme des centres de galaxies.

4.3 Construction de la carte de proposition par contrôle du FDR

Dans le cas du cube HDFS, le cube de variance est relativement fiable, il a donc été utilisé pour réduire les données avant le filtrage adapté. Les données ainsi réduites, sont distribuées

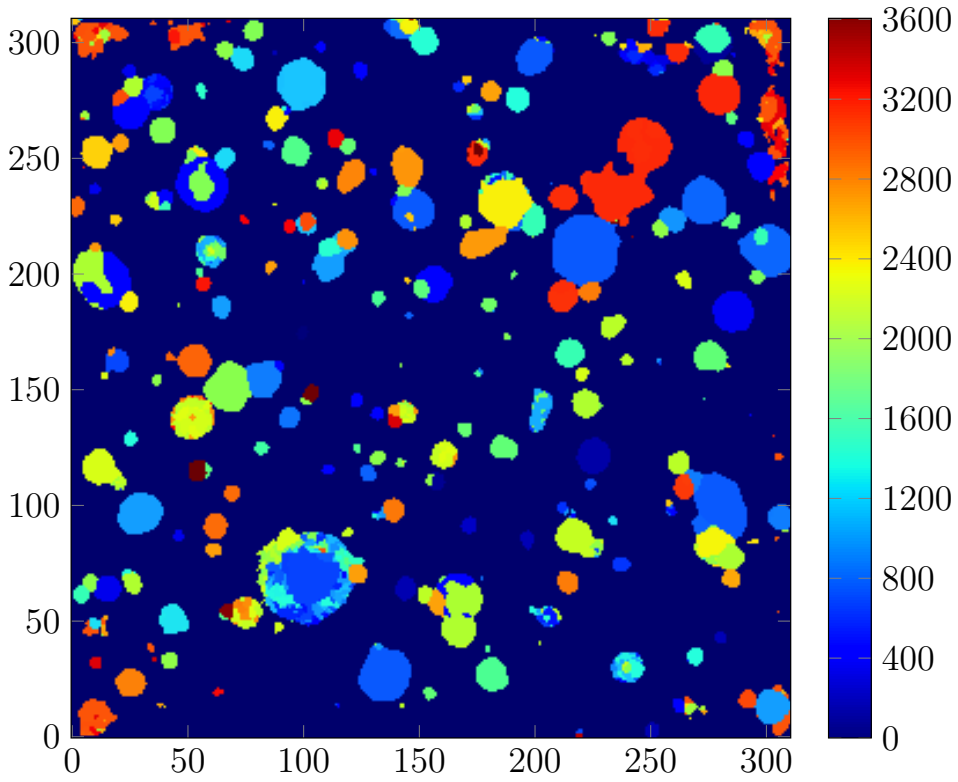


FIGURE 4.9 – Carte des longueurs d'onde indiquant la position de la valeur maximale des spectres des pixels de la carte de proposition.

selon une loi de Student avec un nombre de degré de liberté $N_{poses} - 1$, avec $N_{poses} \simeq 50$ [5]. Les données sont ensuite centrées-réduites avant et après le filtre adapté selon la procédure décrite dans les paragraphes 4.1.3 et 4.1.4. Avant le filtrage adapté, il serait possible d'utiliser la procédure de seuillage par contrôle du FDR de Benjamini-Hochberg puisque les données studentisées constituent l'un des cas étudiés par Benjamini and Yekutieli [2001] pour lequel le contrôle du FDR est maintenu, même pour des données corrélées (à condition que le taux de contrôle du FDR soit inférieur à 50%). En sortie du filtrage adapté, il est difficile de garantir que les pixels du cube filtré respectent la condition PRDS. Cependant, nous pouvons utiliser l'approximation de la loi de Student à $N_{poses} - 1 \sim 50$ degrés de liberté par une loi gaussienne. Dans ce cas, après filtrage adapté, les données restent asymptotiquement gaussiennes multivariées, de matrice de covariance à coefficients non négatifs, la condition PRDS est respectée.

4.3.1 Calcul des p-valeurs

Les données en sortie du filtre adapté, centrées-réduites par la moyenne et la variance estimées par σ-clipping par point fixe, sous l'hypothèse $\mathcal{H}_0$, sont distribuées selon une loi normale. Cette approximation de la loi des données sous $\mathcal{H}_0$ par une loi normale est validée par le diagramme quantiles-quantiles présenté sur la figure 4.5c. Les p-valeurs correspondant aux données HDFS filtrées, centrées-réduites, sont calculées à l'aide de la fonction de répartition de la loi normale. Sur la figure 4.10 nous affichons les p-valeurs, triées par l'ordre croissant, en fonction de leur rang normalisé. Seules les p-valeurs inférieures à $\frac{1}{2}$ sont représentées, puisque nous nous intéressons

5. Le nombre de poses varie d'un pixel à l'autre à cause de l'étape de σ-clipping visant à éliminer les pixels des poses individuelles contaminés par les rayons cosmiques.

aux p-valeurs inférieures à $q \leqslant \frac{1}{2}$ dans la procédure de Benjamini-Hochberg. Sur la courbe des p-valeurs, nous observons deux comportements distincts : des p-valeurs très proches de zéro qui correspondent aux pixels contenant la contribution des sources présentes dans les données, suivies d'une évolution linéaire (pour les p-valeurs comprises entre 0.05 et 0.5) correspondant aux données distribuées sous $\mathcal{H}_0$.

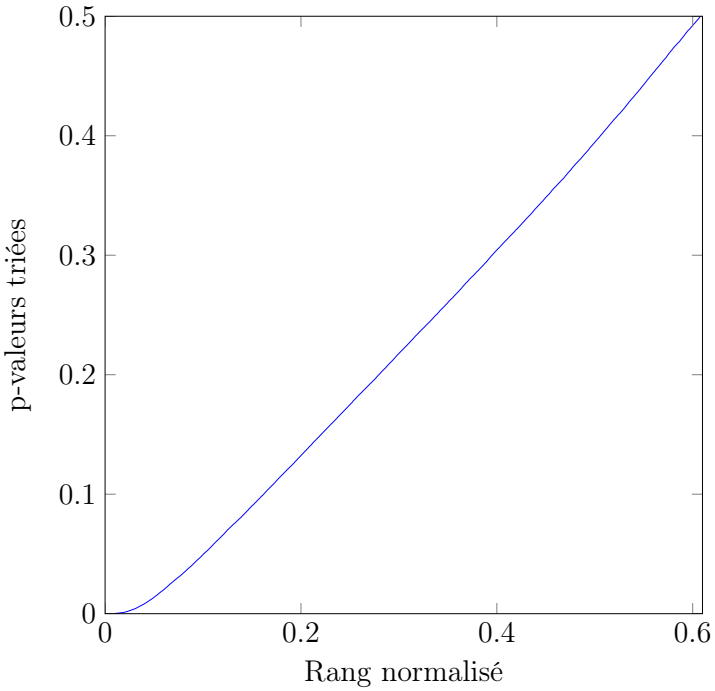
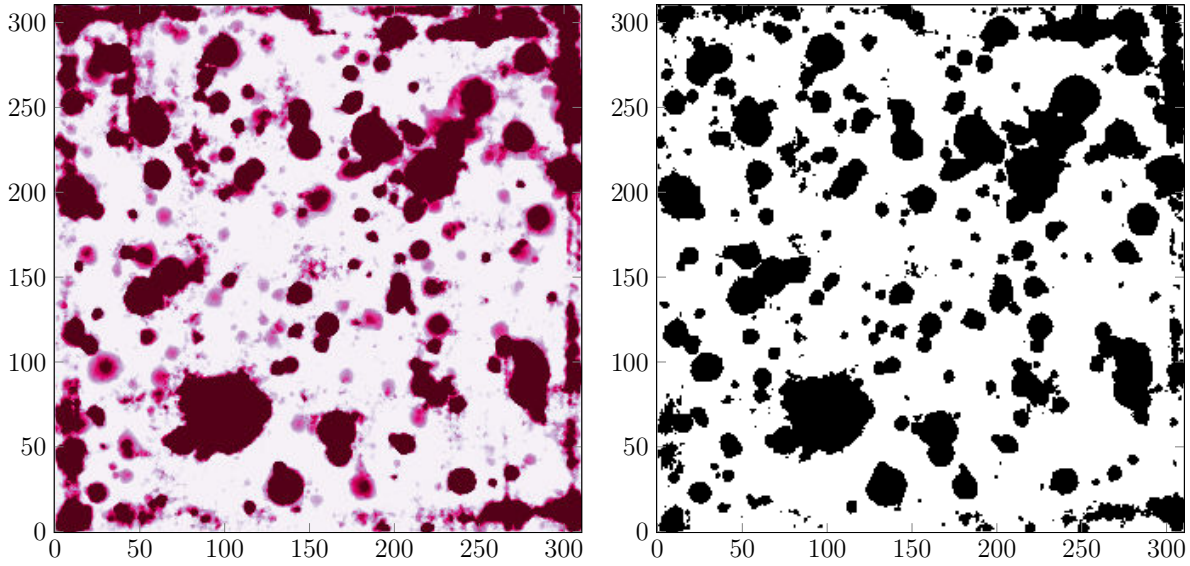


FIGURE 4.10 – Distribution des p-valeurs calculées sur le cube HDFS filtré après centrage et réduction.

4.3.2 Seuillage par la procédure de Benjamini-Hochberg

Le seuillage par la procédure de Benjamini-Hochberg est assez sensible à la présence des structures verticales et horizontales dans le bruit, notamment sur les bords du cube. Le résultat du seuillage par contrôle du FDR à $q = 1\%$ est présenté sur la figure 4.11a, où le cube seuillé a été moyenné selon l'axe des longueurs d'onde. Afin d'éliminer de la carte de proposition les pixels isolés (spectralement) classé dans $\mathcal{C}_1$, les spectres ne présentant pas plus de 5 éléments consécutifs classés sous $\mathcal{H}_1$ sont rejetés de la carte de proposition. Le résultat de ce deuxième seuillage est présenté sur la figure 4.11b. Cette carte ressemble beaucoup à celle obtenue grâce au seuillage par le max-test présenté sur la figure 4.8. Pour les raisons exposées dans le paragraphe 3.4.3.3, il n'est pas possible d'utiliser tous les pixels de la classe $\mathcal{C}_1$ (en noirs sur la figure 4.11b) comme candidats pour être des centres de galaxie.

Afin d'affiner la carte de proposition, nous nous intéressons à la position spectrale de la plus petite p-valeur de chaque spectre conservé dans la carte binaire présentée sur la figure 4.11b. La figure 4.12 présente la carte des longueurs d'onde correspondantes. Nous constatons que les sources qui contiennent principalement une composante continue ne sont pas spectralement homogènes, la statistique de la plus petite p-valeur du spectre n'étant évidemment pas adaptée à ces objets. En revanche les sources dont le spectre contient uniquement une raie d'émission, ont un support spectralement homogène sur la carte figure 4.12.



(a) Projection du cube seuillé (FDR = 1%) (b) Réduction de la carte binaire (FDR = 1%)

Figure 4.11 – Résultats du seuillage du cube HDFS par la procédure de contrôle du FDR par la procédure de Benjamini and Hochberg [1995] dans le cas de tests dépendants. La figure (a) est la carte obtenue en sommant les pixels (binaires) de chaque spectre. Les pixels blancs signifient que tous les pixels du spectre ont été classés dans $\mathcal{C}_0$, plus la couleur est foncée, plus le spectre contient de pixels classés sous $\mathcal{H}_1$. La figure (b) tient compte de l'hypothèse que les raies Lyα présentent une certaine largeur spectrale (quelques bandes consécutives avec la résolution de l'instrument MUSE). Tous les spectres qui ne présentent pas plus de 5 éléments consécutifs classés sous $\mathcal{H}_1$ sont rejetés de la carte de proposition.

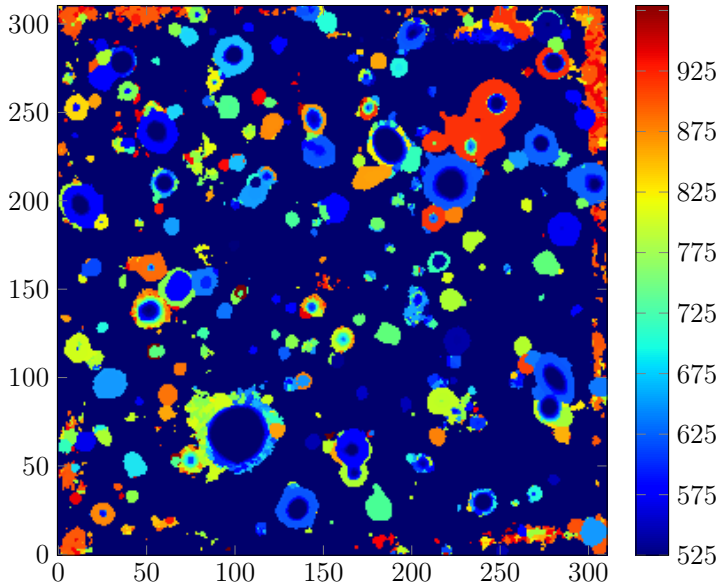


Figure 4.12 – Carte des longueurs d'onde de la plus petite p-valeur de chaque spectre du cube après filtrage adapté.

4.3.3 Perspectives d'amélioration

Tout d'abord, nous pensons qu'une fois les structures verticales et horizontales présentes dans les données atténuées au maximum, le seuillage du cube en trois dimensions par la procédure de Benjamini-Hochberg sera beaucoup plus efficace.

Si le seuillage par contrôle du FDR donne des résultats assez similaires à ceux obtenus par le max-test (mis à part la présence plus accentuées des structures horizontales et verticales), il est difficile d'exploiter ces résultats pour dresser une liste de pixels à proposer lors des mouvements de naissance. Nous avons exploré différentes pistes pour dresser cette liste de pixels candidats, mais elles ne permettent pas d'aboutir à un résultat satisfaisant :

— utiliser les centroïdes 3D des ensembles de pixels du cube seuillé afin de prendre directement en compte le voisinage spatial et en longueur d'onde,
— utiliser les maxima locaux 2D après avoir moyenné le cube seuillé selon l'axe des longueurs d'onde.

Concernant le premier point, l'approche proposée devrait permettre, en particulier, de détecter les maxima locaux 3D dus aux sources dont le spectre contient une unique raie d'émission de faible intensité qui sont situées dans la périphérie (spatiale) des sources spatialement étendue et brillante[6]. En contre-partie, nous risquons d'obtenir plusieurs maxima locaux 3D à différentes longueurs d'onde pour une même source. Les positions de ces différents maxima locaux peuvent même varier dans un petit voisinage spatial autour de la vraie position de la source considérée. Cependant, compte tenu du fait qu'une étape de détection des sources à spectre continu est réalisée avant de lancer la détection sur le cube complet, et qu'un terme de pénalisation des recouvrements entre objets est inclus dans le processus d'acceptation des naissances, cette solution peut être envisageable. Il faudra tout de même attendre une version du cube HDFS où les structures qui quadrillent le cube de données soit mieux atténuées afin de ne pas conserver dans le seuillage trop de contributions de ces structures.

4.4 Détection des objets à spectre continu

Effectuer la détection des sources à spectre continu, ou à raie d'émission suffisamment intense et spectralement étendue[7], sur l'image blanche (voir figure 4.13) présente un intérêt double :

— cela permet de réduire le temps de calcul par rapport au cas où la détection de ces sources aurait lieu directement sur le cube complet,
— dans le cas où certaines sources ne posséderaient qu'une composante spectrale continue de faible intensité, elles pourraient ne pas être mises en évidence par une stratégie de type max-test (développée pour les sources avec une raie d'émission), et donc ne pas être visibles sur la carte de proposition.

4.4.1 Construction de la carte de proposition

Le cube de données (sans traitements spécifiques) est tout d'abord moyenné en longueur d'onde afin d'obtenir l'image blanche. Le bruit étant supposé gaussien dans le cube de données, le bruit présent sur l'image blanche est gaussien. Cette image est ensuite centrée réduite à l'aide de la moyenne et de la variance estimée par σ-clipping par point fixe. Afin de construire la carte de proposition, nous testons chaque pixel $\mathbf{Y}(p,q,\lambda)$ pour les hypothèses suivantes :

$$\begin{cases} \mathcal{H}_0 & : \quad \mathbf{Y}(p,q,\lambda) \sim \mathcal{N}(0,1) \\ \mathcal{H}_1 & : \quad \mathbf{Y}(p,q,\lambda) \sim \mathcal{N}(\mathbf{a},1) \end{cases}$$

6. C'est par exemple le cas des sources #2 et #16 sur le cube DryRun.
7. Il faut qu'une telle raie d'émission subsiste après le moyennage de plus de 3600 bandes spectrales.

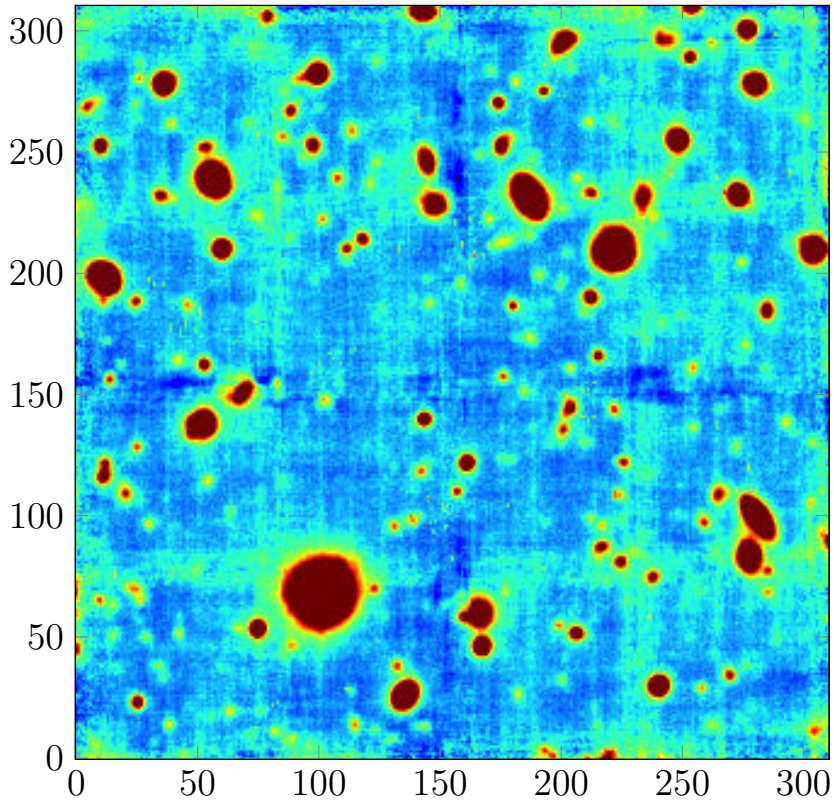


FIGURE 4.13 – Représentation de l'image blanche du cube HDFS.

avec **a** une constante strictement positive représentant la contribution d'une source. Le test permettant d'effectuer la classification binaire : $\mathcal{C}_0$ (bruit) *vs.* $\mathcal{C}_1$ (signal) peut se formuler de la façon suivante :

$$\mathbf{Y}(p,q,\lambda) \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\lessgtr}} \eta, \tag{4.3}$$

où η se déduit à l'aide de la fonction de répartition Φ de la loi normale : $\Phi(\eta) = 1 - p_{FA}$.

Puisque les objets détectés sur l'image blanche ne seront pas modifiés lors de l'échantillonnage de la configuration d'objets globale sur le cube complet, il est dans notre intérêt de ne pas introduire trop d'erreurs lors de la détection sur l'image blanche. Le seuil η est donc fixé à la valeur 3.72 qui correspond à une probabilité de fausse alarme de 0.01% pour le test défini par l'équation (4.3).

4.4.2 Détection à l'aide de l'algorithme d'échantillonnage RJMCMC

La détection sur l'image blanche à l'aide l'algorithme RJMCMC n'est définitivement pas la méthode la plus rapide, mais elle offre l'avantage de renvoyer en sortie un catalogue d'objets pré-formatés pour la détection sur le cube complet. En effet, les objets auront toutes les caractéristiques nécessaires à l'initialisation de la configuration d'objets et à l'estimation finale des spectres de tous les objets (détectés sur l'image blanche et sur le cube complet). Le tableau 4.2 résume les performances de la méthode de détection RJMCMC. La détection sur l'image blanche est très dynamique puisqu'au moins la moitié des mouvements proposés au cours de l'échantillonnage RJMCMC est acceptée. Nous remarquons que les ratios d'acceptation des mouvements de translation et de rotation sur une réalisation de la chaîne RJMCMC sont très élevés, ceci est principalement dû au fait que la modélisation des galaxies spatialement étendues (et donc proches)

par des profils Sersic à supports elliptiques est trop simple. Il faut alors un grand nombre de mouvements sur les objets correspondants avant de trouver la meilleure représentation possible des contributions de ces sources. Le nombre d'objets de la configuration fluctue beaucoup au cours du processus d'échantillonnage, cependant, la configuration d'objets finit par se stabiliser à l'itération 12798 (sur 15601 itérations). Il faut noter que sur différentes réalisations de la chaîne RJMCMC, le nombre d'itérations nécessaires à la convergence du nombre d'objets dans la configuration peut varier entre 10000 et 40000 pour les cas les plus extrêmes. Cette durée est assez sensible à l'ordre dans lequel sont échantillonnés les objets de la configuration. Les objets présentant les intensités les moins élevées ne se stabilisent pas dans la configuration tant que les objets les plus brillants n'ont pas été détectés. Ceci s'explique par le fait que tant que les objets brillants n'ont pas été correctement détectés, l'estimation de la variance du bruit est biaisée par les fort résidus, ce qui entraîne une grande variabilité de la configuration d'objets.

Ratio de naissances acceptées/proposées	54.7%
Ratio de morts acceptées/proposées	48.2%
Ratio de translations acceptées/proposées	72.2%
Ratio de rotations acceptées/proposées HST	84.5%
Ratio de modifications acceptées/proposées	55.1%

TABLEAU 4.2 – Description de l'algorithme d'échantillonnage RJMCMC avec les ratios d'acceptation des différents mouvements sur la configuration d'objets lors de la détection sur l'image blanche.

La détection sur l'image blanche est représentée sur la figure 4.14. Au total 127 sources ont été détectées sur cette image. L'étoile la plus brillante a été initialisée manuellement car son profil d'intensité spatial, qui permet d'estimer la FSF de l'instrument pour ce cube de données à chaque longueur d'onde λ, est très différent des profils Sersic proposés par l'algorithme. Puisque sa position et son profil d'intensité sont connus, il est possible d'initialiser la configuration avec cette étoile, cela permet d'éviter notamment des surdétections au niveau de l'étoile. Parmi les 127 objets détectés, 120 appartiennent au catalogue HST (dont 84 recensées également sur le cube MUSE). Les 7 objets n'étant pas répertoriés dans le catalogue sont pour quatre d'entre eux des fausses détections localisées sur des structures horizontales et verticales, là où la variance est très grande. Les trois autres objets sont localisés dans la périphérie de l'étoile la plus brillante (en bas à gauche de l'image) à proximité d'autres sources et constituent des fausses détections. Finalement le catalogue construit à l'issu de la détection contient 5.5% de fausses découvertes. Cette proportion est relativement stable lorsque plusieurs chaines de détection RJMCMC sont lancées en parallèle sur l'image blanche. Le nombre d'objets total varie entre 122 et 133 et les sources ne coïncidant pas avec le catalogue HST sont en général les mêmes que celles du résultat présenté sur la figure 4.14.

4.5 Détection des objets à raies d'émissions

L'algorithme a été principalement conçu pour détecter des sources quasi-ponctuelles dans les trois dimensions du cube. La grande majorité des sources à spectre continu ont été détectées lors de la phase de détection sur l'image blanche, permettant ainsi à l'algorithme de se focaliser sur la recherche de sources dont le spectre comporte principalement une raie d'émission.

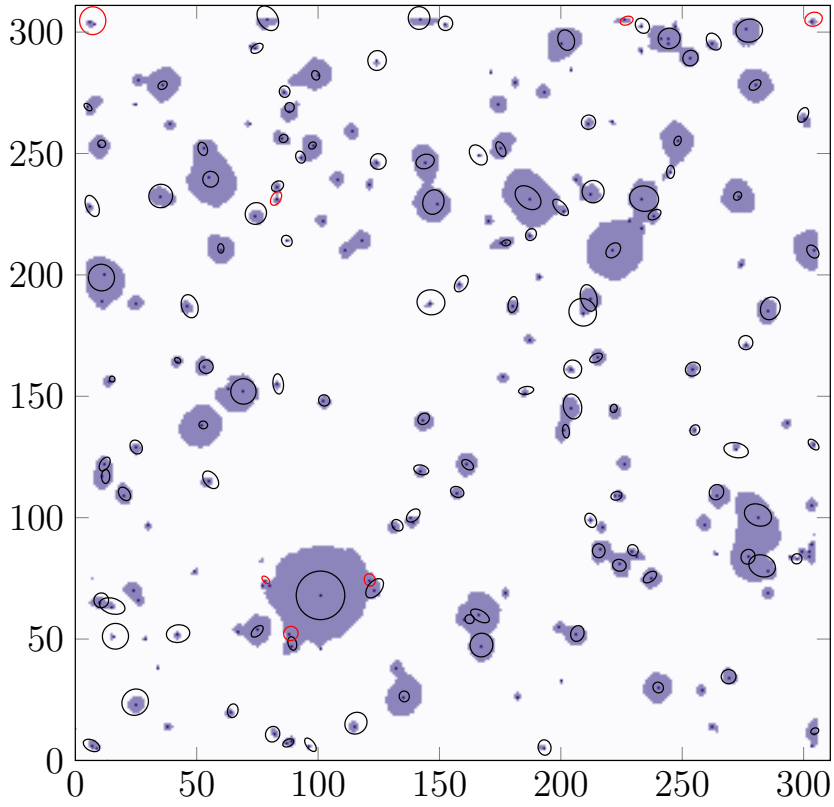


FIGURE 4.14 – Résultats de la détection des sources sur l'image blanche obtenue en moyennant les 3641 images formant le cube HDFS. L'image de fond correspond à la carte de proposition utilisée pour effectuer la détection. Les ellipses modélisent le support des objets détectés à l'aide de la carte de proposition (image de fond) calculée sur l'image blanche. Les ellipses de couleurs noires sont des objets qui correspondent à des sources répertoriées dans le catalogue HST et les ellipses rouges sont des erreurs de détection.

4.5.1 Résultat de la détection

Pour établir ce résultat, nous utilisons la carte de proposition obtenue en appliquant le max-test avec une probabilité de fausse alarme de $p_{FA} = 0.1\%$, voir la carte présentée sur la figure 4.8. Le résultat de la détection sur le cube de données hyperspectrales est présenté sur la figure 4.15. En plus des 127 sources détectées sur l'image blanche (ellipses noires sur la figure 4.15), 171 objets ont été ajoutés à la configuration estimée au sens du maximum *a posteriori* (ellipses fushia sur la figure 4.15).

4.5.2 Convergence de l'algorithme

La détection sur le cube complet a été réalisée en 16424 itérations, la configuration d'objets qui maximise la densité *a posteriori* a été échantillonnée à l'itération 16415 tandis que le nombre d'objets est stable depuis l'itération 10404. Sur la figure 4.16, nous représentons l'évolution du nombre d'objets dans la configuration d'objets qui maximise la densité *a posteriori* au fur et à mesure des itérations. Contrairement à la détection sur l'image blanche, le nombre d'objets ne diminue pas durant le processus d'échantillonnage. L'intégralité des mouvements de mort proposés est rejetée lors de l'étape d'acceptation-rejet de la procédure introduite dans l'encadré 3.3 dans le chapitre 2 car il n'y a pas de régularisation dans le modèle Bayésien sur les intensités.

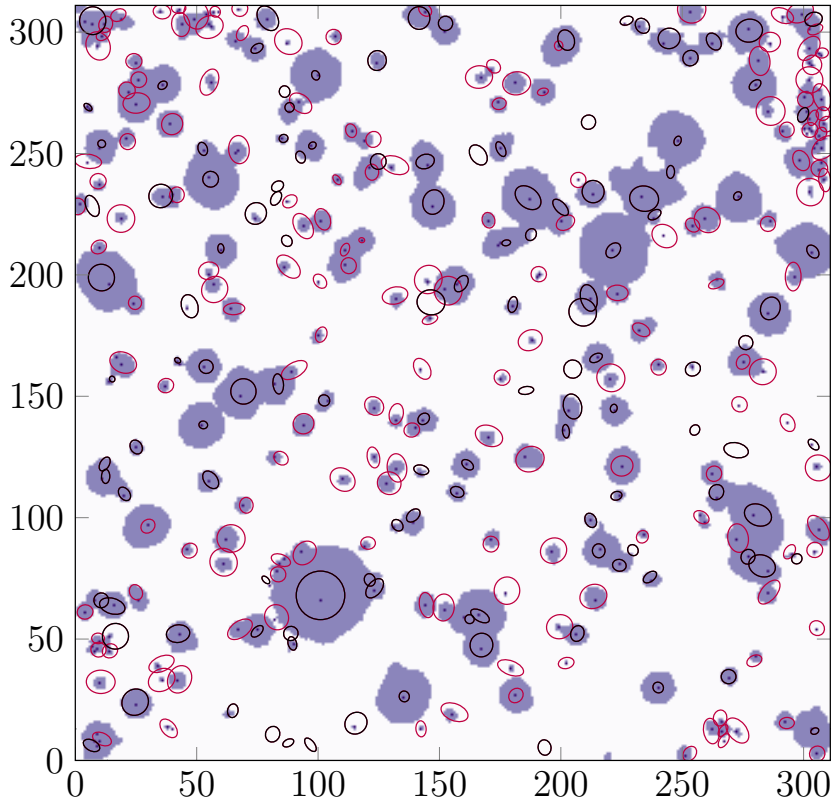


FIGURE 4.15 – Résultats de la détection des sources sur le cube HDFS. Les ellipses noires modélisent le support des objets pré-détectés sur l'image blanche. Les ellipses fushia modélisent le support des objets détectés sur le cube à l'aide de la carte de proposition obtenue par le max-test (image de fond).

Le contrôle est uniquement assuré par la fonction d'intensité du processus ponctuel calculée à partir des prétraitements (carte de proposition obtenue par le max-test ou par contrôle du FDR selon le prétraitement choisi)..

La dynamique de l'échantillonnage RJMCMC correspondant au résultat de la détection sur le cube HDFS présenté dans ce chapitre est résumée dans le tableau 4.3.

Ratio de naissances acceptées/proposées	6.9%
Ratio de morts acceptées/proposées	0 %
Ratio de translations acceptées/proposées	23.1%
Ratio de rotations acceptées/proposées HST	40.6%
Ratio de modifications acceptées/proposées	16.9%

TABLEAU 4.3 – Description de l'algorithme d'échantillonnage RJMCMC avec les ratios d'acceptation des différents mouvements sur la configuration d'objets.

Le ratio de naissances acceptées est très faible, ceci peut s'explique par plusieurs points :

— la carte de proposition possède un nombre limité de centres à proposer lors des mouvements de naissance, une fois tous les objets "probables" au sens du maximum *a posteriori* acceptés, les autres pixels n'étant pas retenus comme centre d'objets sont proposés à nouveau jusqu'à la fin de l'algorithme,

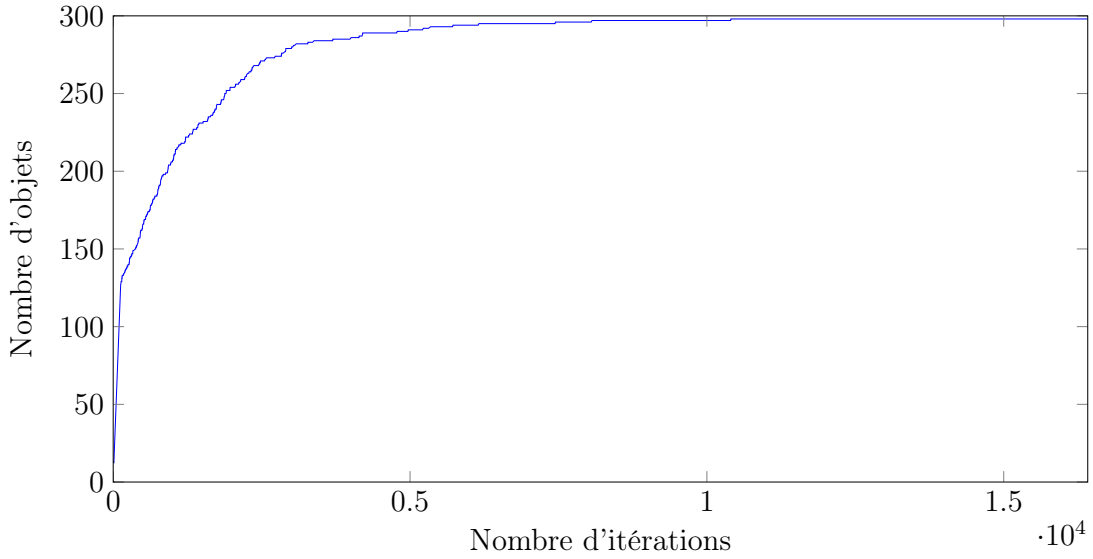


FIGURE 4.16 – Evolution du nombre d'objets dans la configuration estimée au sens du maximum *a posteriori* au fur et à mesure des itérations.

— aucune mort n'est acceptée à cause du terme d'attache aux données très fort lorsqu'un objet recouvre spatialement une source du cube,
— si la carte de proposition contient des pixels de bruit supérieurs au seuil η et identifiés comme des maxima locaux, les objets proposés sur ces pixels ne seront quasiment jamais retenus car le terme d'attache aux données n'est pas suffisamment fort pour faire accepter la naissance. En revanche lorsqu'un objet est accepté, le terme d'attache aux données est en général suffisamment fort pour empêcher la mort.

Ces trois points limitent forcément le ratio de naissances acceptées par rapport au nombre de naissances proposées.

Notons enfin que les paramètres de variances σ_λ, $\lambda = 1 \cdots \Lambda$, introduits dans le modèle bayésien (cf. chapitre 2) jouent le rôle d'une température dans l'algorithme d'échantillonnage. La variance baisse avec la convergence d'algorithme (les données sont de mieux en mieux expliquées par la configuration d'objets) ce qui entraîne la convergence vers la solution optimale.

4.5.3 Estimation des paramètres de moyenne et variance du bruit

Les paramètres de moyenne et de variance du bruit pour chacune des Λ longueurs d'onde du cube sont échantillonnés par l'algorithme de Gibbs présenté dans l'encadré 2.1 au chapitre 2. Les estimateurs au sens du maximum *a posteriori* sont présentés sur les figures 4.17 et 4.18. Nous comparons ces estimations à celles fournies par la méthode de σ-clipping par point fixe appliquée à chaque image du cube HDFS (données brutes sans prétraitement). L'estimation au sens du maximum *a posteriori* est légèrement plus élevée, notamment au niveau des basses longueurs d'onde ($\lambda \leqslant 550$nm), mais les variations des deux estimateurs sont assez similaires. La variance estimée au sens du maximum *a posteriori* est plus élevée que l'estimation par σ-clipping par point fixe. Deux raisons peuvent expliquer cette surestimation de la variance par l'approche bayésienne :

— la densité conditionnelle *a posteriori* du paramètre de variance dépend de la configuration d'objets détectés (voir équation (2.27)). Si la configuration d'objets estimée au sens du maximum *a posteriori* ne modélise pas tous les objets présents dans l'image et s'il reste des résidus dus aux erreurs de modélisation des sources, l'estimation de la variance sera

systématiquement plus élevée que la variance réelle du bruit,

— l'estimation par σ-clipping par point fixe est obtenue sur les données tronquées afin de s'affranchir des pixels contenant une contribution positive, cela permet notamment de ne pas tenir compte des pixels situés sur les structures verticales et horizontales du bruit et sur les bords du cube où la variance est plus grande. En revanche, l'estimation au sens du maximum *a posteriori* n'exclue pas ces pixels, la variance estimée est donc biaisée.

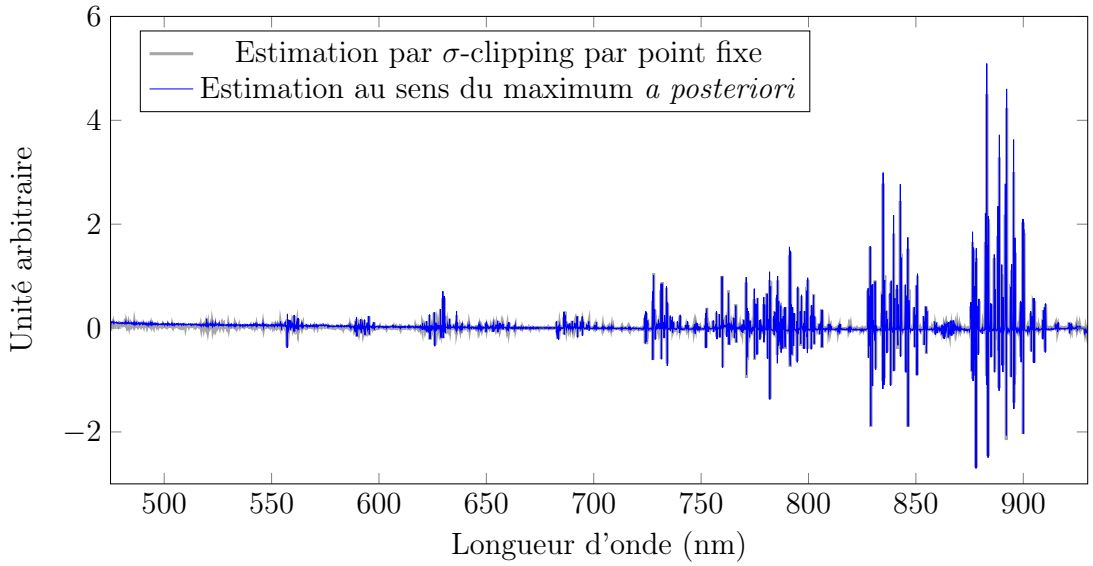


FIGURE 4.17 – Estimation au sens du maximum *a posteriori* de la moyenne du bruit (courbe bleue) superposée à l'estimation par σ-clipping par point fixe sur le cube de données HDFS (courbe grise).

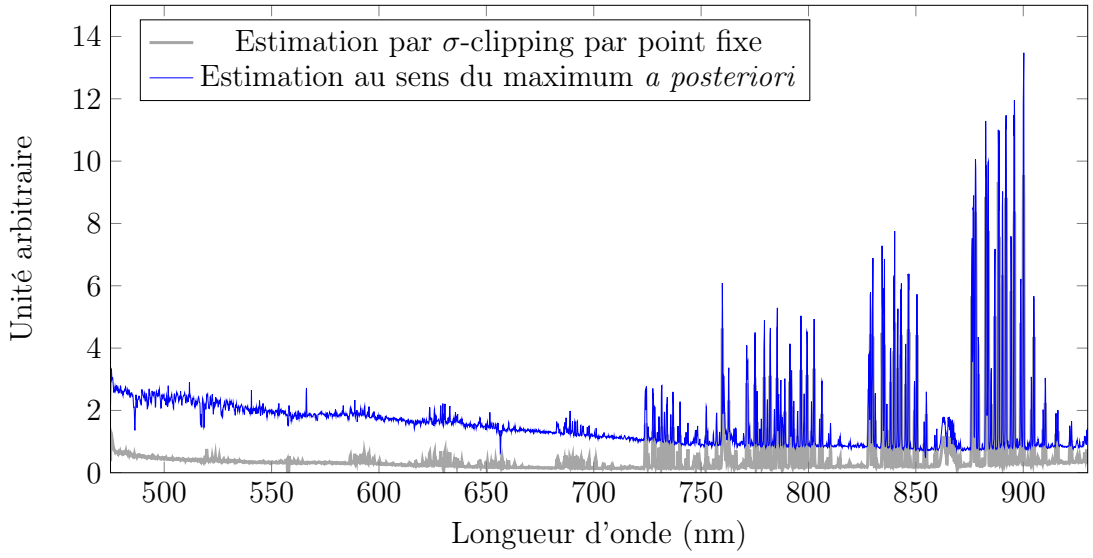


FIGURE 4.18 – Estimation au sens du maximum *a posteriori* de la variance du bruit (courbe bleue) superposée à l'estimation par σ-clipping par point fixe sur le cube de données HDFS (courbe grise).

4.6 Analyse des résultats

Nous allons maintenant comparer le catalogue d'objets détectés à l'aide de notre algorithme avec le catalogue des sources répertoriées à partir de l'image HST et avec le catalogue des sources de l'image HST qui sont visibles dans le cube MUSE.

4.6.1 Comparaison avec les catalogues HST et MUSE

A l'aide du logiciel TOPCAT (*Tool for operations on catalogues and tables*) qui permet de comparer des catalogues d'objets astrophysiques, nous établissons les performances de l'algorithme de détection proposé dans ces travaux dans le tableau 4.4.

Nombre d'objets détectés sur l'image blanche	127
Nombre d'objets à raie d'émission détectés sur le cube complet	171
Nombre total d'objets détectés	298
Nombre d'objets correspondant au catalogue MUSE	166
Nombre d'objets correspondant au catalogue HST	78
Nombre d'objets qui ne sont pas dans les catalogues	54
... dont potentielles galaxies	6

Tableau 4.4 – Résultats de la détection sur le cube HDFS.

Les sources recensées dans le catalogue HST sont représentées par tous les cercles, à l'exception des verts, sur la figure 4.19. Les cercles oranges et rouges modélisent les 189 sources du catalogue HST que les astrophysiciens ont détectées et mesurées [8] lors de l'exploration semi-manuelle du cube HDFS acquis par MUSE. Ces 189 sources extraites du catalogue HST sont les sources du catalogue MUSE dont la description a été donnée dans le tableau 1.2 au chapitre 1. Parmi les 189 sources, 7 sont superposées à d'autres galaxies, elles ont pu être détectées séparément sur le cube MUSE grâce à la connaissance *a priori* apportée par la résolution plus fine du HST, mais avec la résolution de MUSE, elles ne sont pas automatiquement séparables de leurs voisines. Le max-test n'offre pas la possibilité de distinguer deux sources spatialement superposées, tout comme notre algorithme de détection n'est de toute façon pas capable de séparer des sources superposées. Il y également 3 sources qui ont été initialement répertoriée dans le catalogue MUSE, mais avec l'amélioration des données, il s'avère qu'il s'agit sans doute de fausses détections dus à des résidus de traitement présents dans la première version du cube qui a servi à construire le catalogue MUSE. Parmi les sources du catalogue MUSE, 3 sources ont un spectre constitué d'une faible composante continue et ont pu être identifiées grâce à leur raie d'absorption, or le max-test est conçu pour détecter les raies d'émission. Finalement en ne tenant pas compte de ces 13 sources problématiques, nous avons détecté 166 sources sur 175, soit un rappel [9] de 94.5%. Seuls 44% des sources du catalogue HST ont été détectées par notre algorithme, mais rappelons que le spectre de la majorité des sources détectées sur l'image HST est constitué principalement d'une composante continue et que l'algorithme, ainsi que les prétraitements mis en oeuvre pour construire la carte de proposition, ne sont pas conçus pour la détection de ce type de galaxie. De plus un grand nombre des galaxies détectées sur l'image HST sont invisibles dans les données MUSE du fait de la résolution spatiale de MUSE beaucoup plus faible que celle du télescope Hubble.

8. Mesure du redshift.
9. Le rappel correspond au nombre de sources détectées appartenant au catalogue MUSE sur le nombre total de sources du catalogue MUSE.

Parmi les 54 sources ne correspondant pas aux catalogues HST et MUSE, 18 objets sont localisés dans le coin supérieur droit de l'image sur une structure verticale de bruit dont la variance est très élevée. La présence de de type de structures à valeurs positives est problématique pour notre algorithme qui acceptera toujours la naissance d'objets dans ces zones afin d'expliquer au mieux la présence d'un signal significativement différent du fond. La plupart de ces structures ont été atténuées lors des prétraitements effectués par les astrophysiciens lors de la construction du cube de données. Des travaux en cours au sein du consortium devraient permettre de fournir dans les mois à venir de nouvelles versions du cube où les structures résiduelles seront encore mieux atténuées. Sans compter les 18 sources situées sur la structure de bruit verticale, et en retirant des 54 objets, les potentielles galaxies, la proportion de fausses découvertes dans la liste des objets retournée par l'algorithme est d'environ 11% (soit une précision de 89%).

4.6.2 Analyses des spectres de potentielles nouvelles galaxies

Nous nous intéressons maintenant aux 6 galaxies potentielles détectées par notre algorithme, désignées par des cercles verts sur l'image 4.19. Ces sources présentent un intérêt puisqu'elles sont spatialement isolées des sources répertoriées dans le catalogue HST et elles sont présentes sur la carte de proposition avec un support de quelques pixels spatialement connectés. Les sources détectées qui ne sont pas dans le catalogue HST sont listées dans l'annexe H. Les identifiants des 6 sources qui nous intéressent sont : #130, #144, #184, #190, #243 et #284. Il n'y a pas de notion d'ordre parmi les sources détectées, ces identifiants permettent seulement de les classer dans un catalogue.

4.6.2.1 Etude de la source #130

La source #130 apparaît sur la carte de proposition de façon isolée (voir figure 4.19) ainsi que sur la carte des valeurs maximales des spectres après filtrage adapté (voir figure 4.20d). Son spectre contient effectivement une raie d'émission à la longueur d'onde $\lambda = 678.375$nm, seulement, en s'intéressant à la carte des longueurs d'onde, figure 4.20e, et au spectre de la galaxie située juste en dessous de la source #130, nous nous apercevons qu'il s'agit de la même source, il s'agit d'une excroissance de la galaxie. Ce cas illustre parfaitement un des types d'erreurs caractéristiques de notre algorithme de détection : les galaxies dont le support spatial est complexe, *i.e.* pour lequel l'approximation elliptique n'est pas adaptée, entraîneront souvent des cas de surédétections. L'objet #130 n'est donc pas une nouvelle galaxie.

Notons que tous les spectres présentés par la suite sont estimés sur le cube de données brut, c'est-à-dire avant centrage et réduction et sans filtrage adapté. Ceci explique le fait que la raie d'émission détectée ne correspond pas à la valeur maximale du spectre.

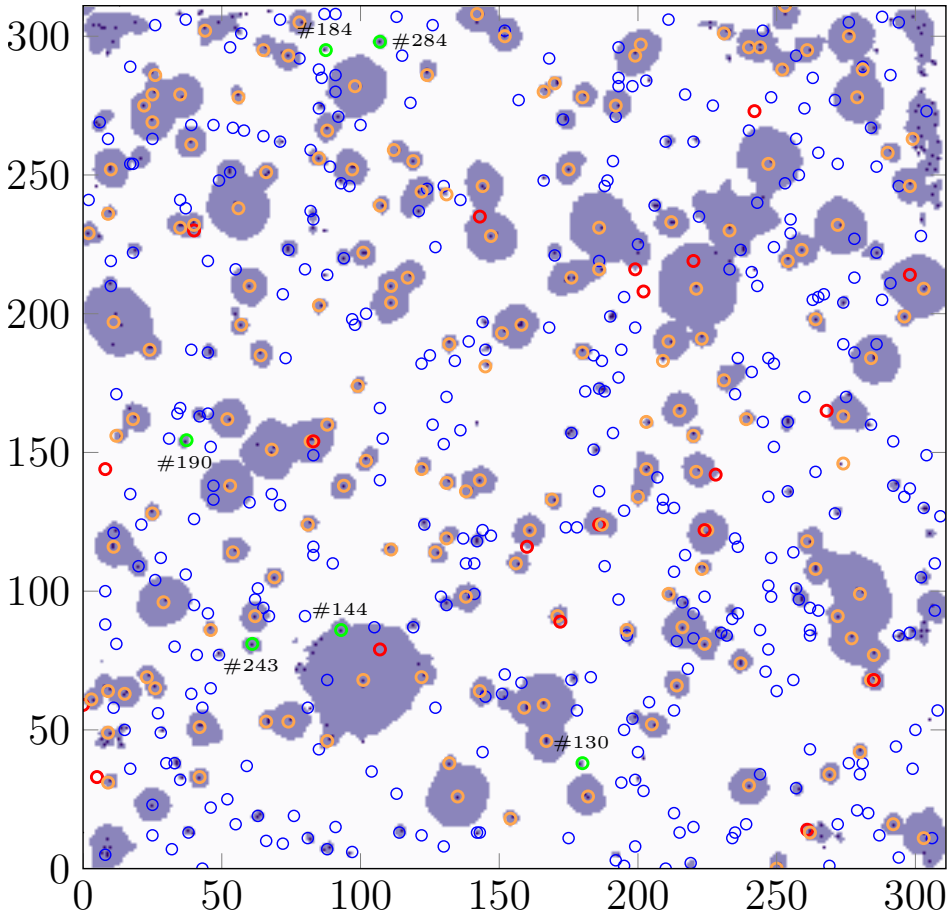
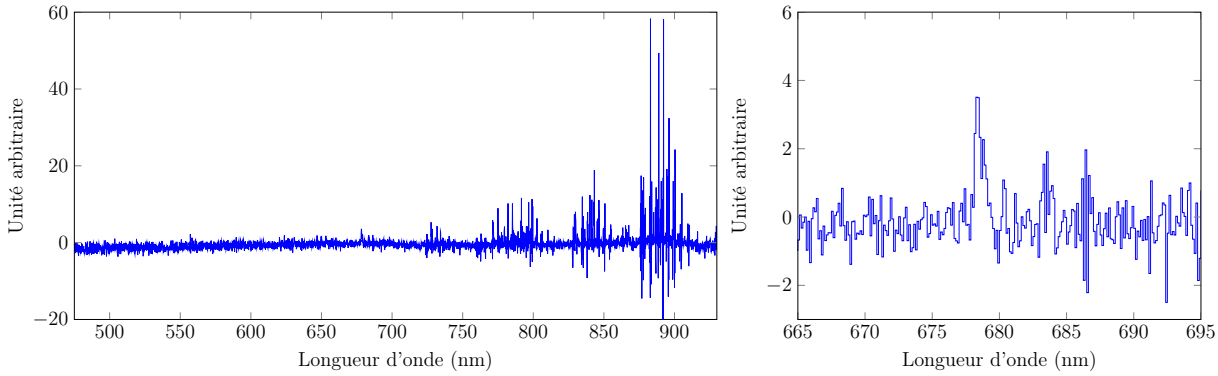


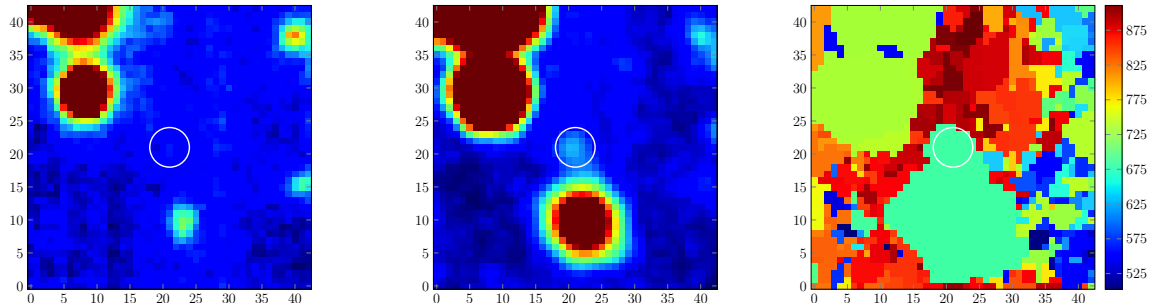
FIGURE 4.19 – Représentation des sources cataloguées à partir de l'image HDFS observée par le télescope spatial Hubble. Les sources sont modélisées par des cercles. Les cercles bleus sont les sources du catalogue HST qui possèdent un spectre exclusivement continu, les cercles oranges permettent de distinguer les sources du catalogue HST qui ont été également observées dans le cube de données MUSE. les cercles rouges représentent les sources du catalogue MUSE qui ne peuvent pas être proposés avec les cartes de proposition utilisées pour la détection sur l'image blanche et sur le cube complet, car il n'y a pas de maxima locaux sur la carte des valeurs maximales des spectres aux emplacements correspondant. L'image de fond est la combinaison de la carte de proposition utilisée pour l'image blanche et la carte de proposition utilisée pour détecter les sources à raie d'émission seule ($p_{FA} = 0.1\%$). Les cercles verts représentent les 6 galaxies potentielles détectées par l'algorithme et qui ne sont pas dans les catalogues MUSE et HST.

4.6.2.2 Etude de la source #144

La source #144 est particulièrement intéressante puisqu'elle se situe en périphérie de l'étoile la plus brillante du champ. La source n'est pas visible sur l'image blanche (figure 4.21c), en revanche sur la carte des valeurs maximales des spectres après filtrage adapté (figure 4.21d), nous pouvons apercevoir un signal de faible amplitude qui ressemble à une excroissance de l'étoile. Cependant, la carte des longueurs d'onde (figure 4.21e) indique que ce signal est homogène en longueur d'onde et que le maximum du spectre est localisé dans une zone différente du spectre que le maximum de l'étoile. Nous observons effectivement dans le spectre estimé de la source #144 une raie d'émission à la longueur d'onde λ =754.875nm (voir figure 4.21b).



(a) Spectre de l'objet #130 (b) Zoom



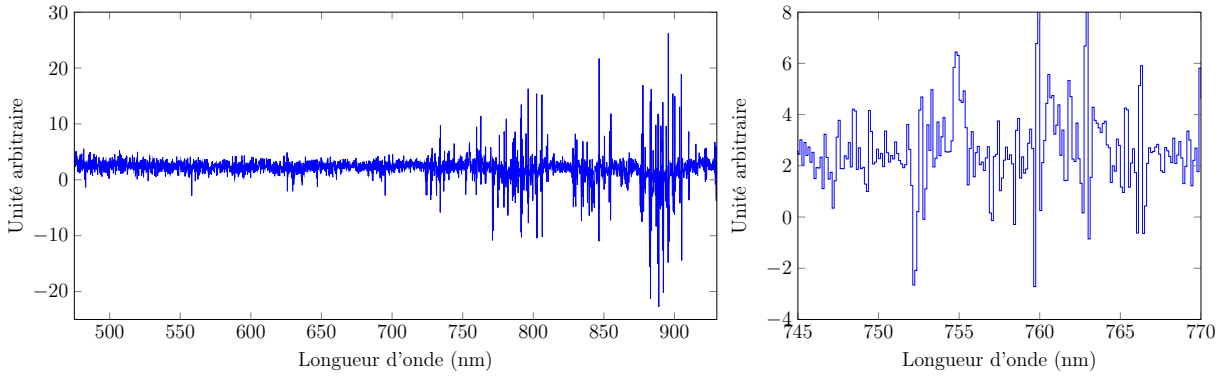
(c) Image blanche (d) Valeurs maximales (e) Longueurs d'onde

FIGURE 4.20 – Analyse de l'objet #130 avec : le spectre estimé correspondant à l'objet #130 (a) et le zoom sur la raie d'émission détectée à l'aide du max-test à la longueur d'onde $\lambda = 678.375$nm (b), le cercle blanc symbolise la position de l'objet #130 sur l'image blanche (c), la carte des valeurs maximale (d) et la carte des longueurs d'onde (e).

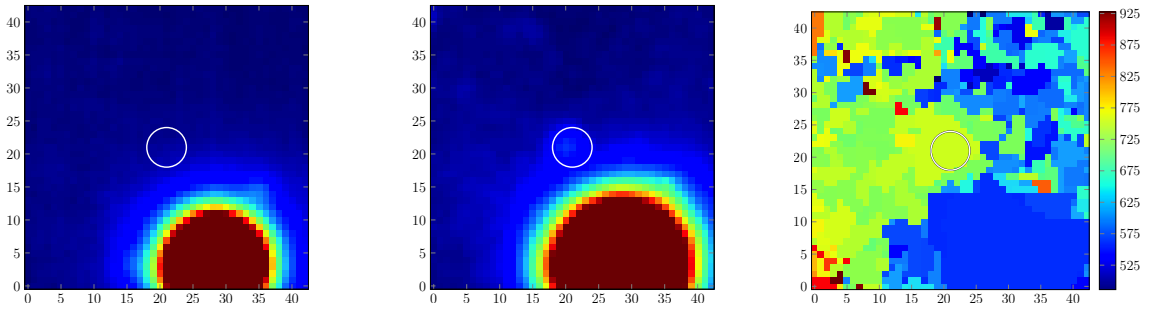
4.6.2.3 Etude des sources #184 et #284

Les sources #184 et #284 sont situées à proximité de la même galaxie brillante, voir figure 4.19. Elles sont toutes les deux de types émetteurs, *i.e.* elles n'existent que sur les bandes spectrales centrées autour de la longueur d'onde de leur raie d'émission. Sur la figure 4.22 sont présentées les images bandes étroites [10] centrées sur les longueurs d'onde respectives des raies d'émission des sources #184 et #284.

10. Une image bande étroite est obtenue en sommant les images du cube sur quelques longueurs d'onde successives (ici 20 bandes spectrales centrées autour de la longueur d'onde de la valeur maximale du spectre de chacune des galaxies).

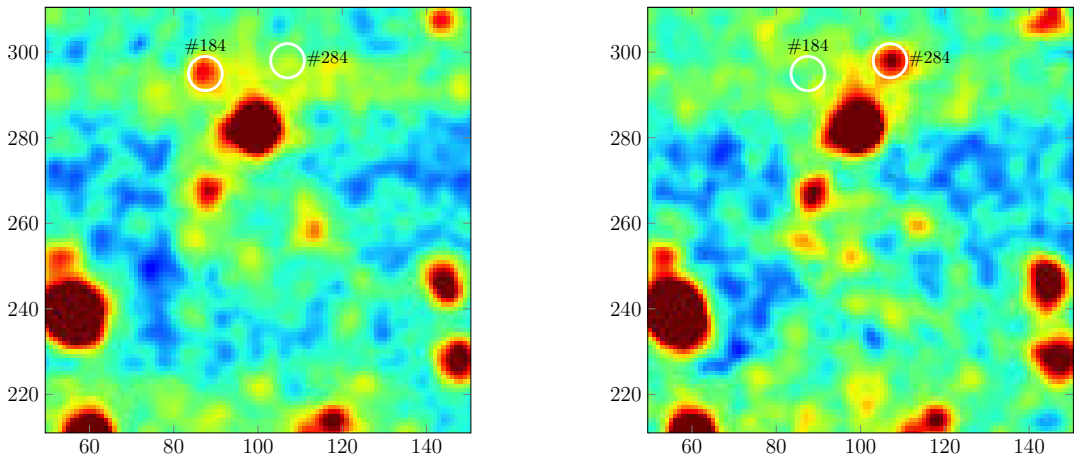


(a) Spectre de l'objet #144 (b) Zoom



(c) Image blanche (d) Valeurs maximales (e) Longueurs d'onde

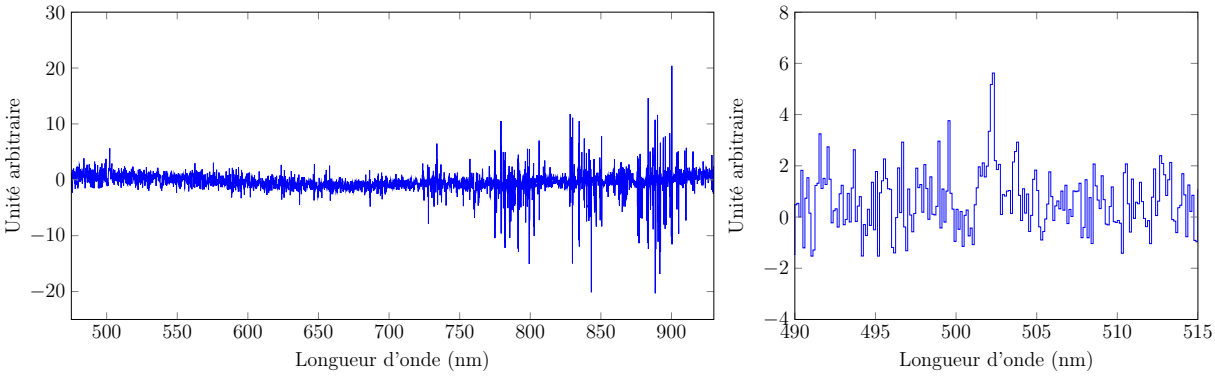
FIGURE 4.21 – Analyse de l'objet #144 avec : le spectre estimé correspondant à l'objet #144 (a) et le zoom sur la raie d'émission détectée à l'aide du max-test à la longueur d'onde $\lambda = 754.875$nm (b), le cercle blanc symbolise la position de l'objet #144 sur l'image blanche (c), la carte des valeurs maximale (d) et la carte des longueurs d'onde (e).



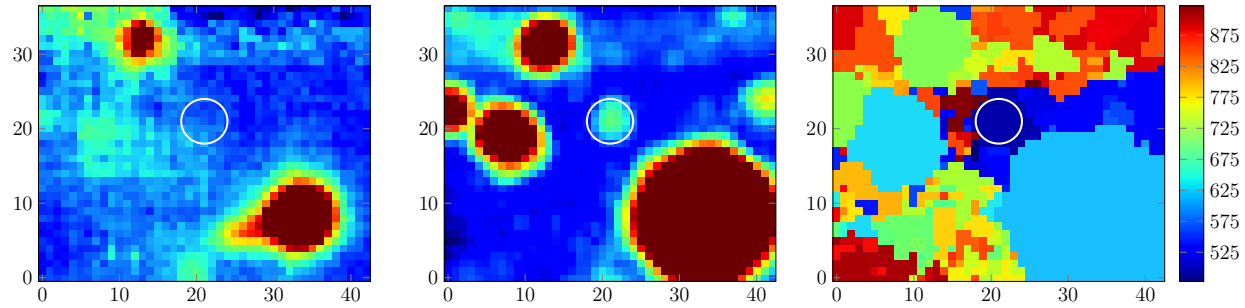
(a) Bande étroite à $\lambda = 502.25$nm (b) Bande étroite à $\lambda = 533$nm

FIGURE 4.22 – Images bande étroite autour des raies d'émission des sources #184 et #284.

Le spectre de la source #184, figures 4.23a et 4.23b est très bruité aux basses longueurs d'onde, cependant la raie d'émission parvient à se détacher suffisamment pour être détectée parmi le bruit. La figure 4.23e montre que le support de la source est spectralement homogène, les maxima des spectres de ce support (spatial) sont localisés à la même longueur d'onde.



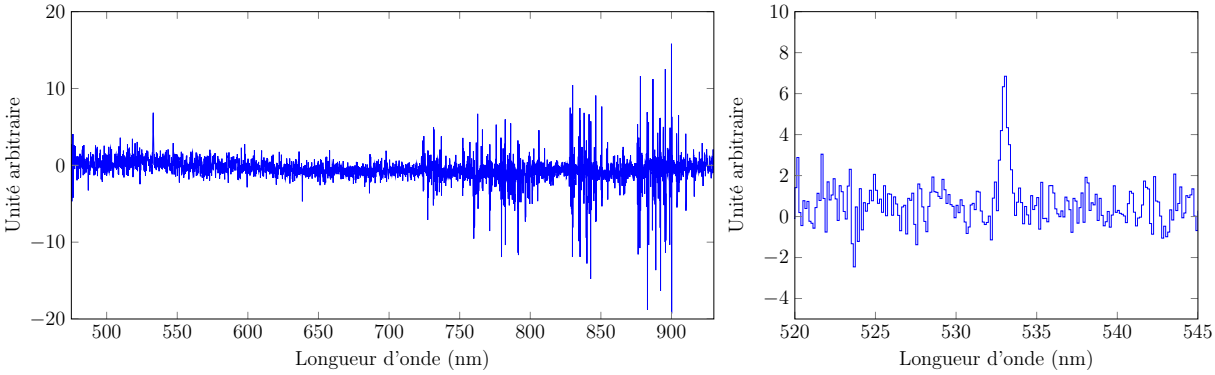
(a) Spectre de l'objet #184 (b) Zoom



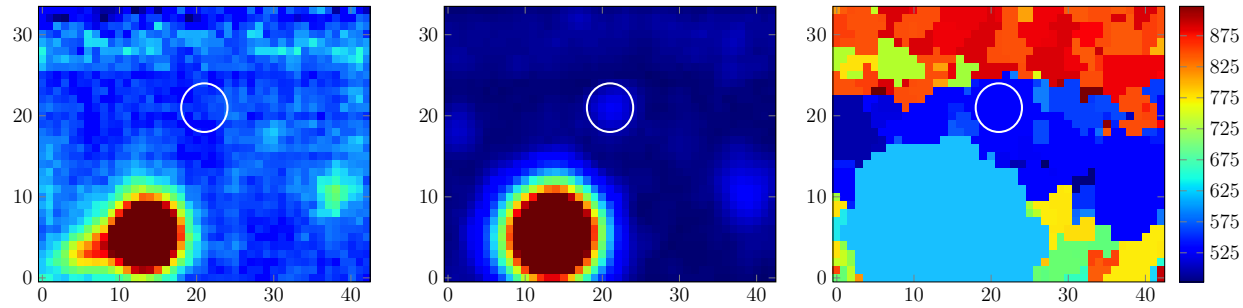
(c) Image blanche (d) Valeurs maximales (e) Longueurs d'onde

FIGURE 4.23 – Analyse de l'objet #184 avec : le spectre estimé correspondant à l'objet #184 (a) et le zoom sur la raie d'émission détectée à l'aide du max-test à la longueur d'onde $\lambda = 502.25$nm (b), le cercle blanc symbolise la position de l'objet #184 sur l'image blanche (c), la carte des valeurs maximale (d) et la carte des longueurs d'onde (e).

L'étude du spectre de la source #284, figures 4.24a et 4.24b, nous révèle la présence d'une raie d'émission de forte intensité centrée à la longueur d'onde λ =533nm tandis que le reste du spectre ne contient que du bruit. A nouveau, nous retrouvons une homogénéité spectrale sur la carte des longueurs d'onde (figure 4.24e).



(a) Spectre de l'objet #284 (b) Zoom

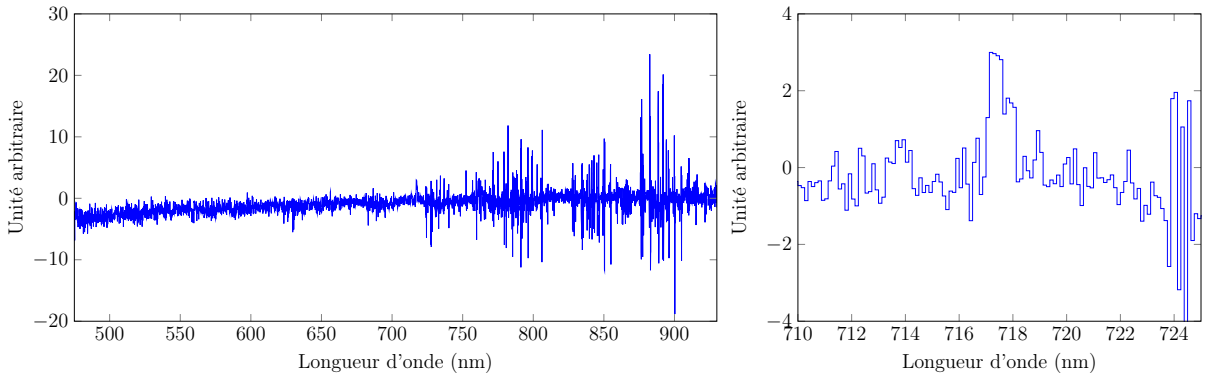


(c) Image blanche (d) Valeurs maximales (e) Longueurs d'onde

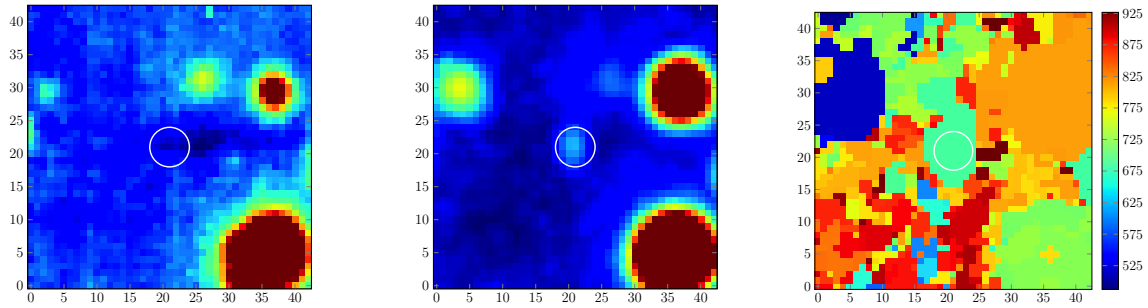
Figure 4.24 – Analyse de l'objet #284 avec : le spectre estimé correspondant à l'objet #284 (a) et le zoom sur la raie d'émission détectée à l'aide du max-test à la longueur d'onde λ =533nm (b), le cercle blanc symbolise la position de l'objet #284 sur l'image blanche (c), la carte des valeurs maximale (d) et la carte des longueurs d'onde (e).

4.6.2.4 Etude de la source #190

La source identifiée par l'objet #190 présente un spectre légèrement négatif sur la première moitié de la gamme de longueurs d'onde. Elle est située au niveau d'une structure horizontale dont l'intensité moyenne est négative (voir sur l'image blanche, figure 4.25c, où les pixels bleu foncé sont des valeurs négatives), ce qui explique les valeurs négatives dans le spectre. Cependant, lors de l'étape de seuillage par le max-test, une raie d'émission a été détectée à $\lambda = 717.375$nm (voir figures 4.25a et 4.25b) et les résultats du max-test (carte des maxima des spectres après filtrage adapté, figure 4.25d, et carte des longueurs d'onde correspondantes, figure 4.25e) montre la présence d'un ensemble de pixels spatialement et spectralement cohérent.



(a) Spectre de l'objet #190 (b) Zoom

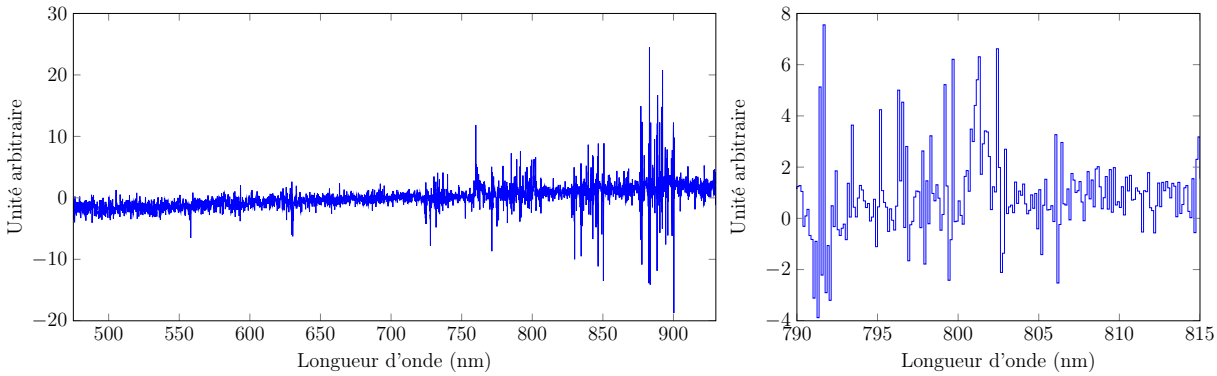


(c) Image blanche (d) Valeurs maximales (e) Longueurs d'onde

FIGURE 4.25 – Analyse de l'objet #190 avec : le spectre estimé correspondant à l'objet #190 (a) et le zoom sur la raie d'émission détectée à l'aide du max-test à la longueur d'onde $\lambda = 717.375$nm (b), le cercle blanc symbolise la position de l'objet #190 sur l'image blanche (c), la carte des valeurs maximale (d) et la carte des longueurs d'onde (e).

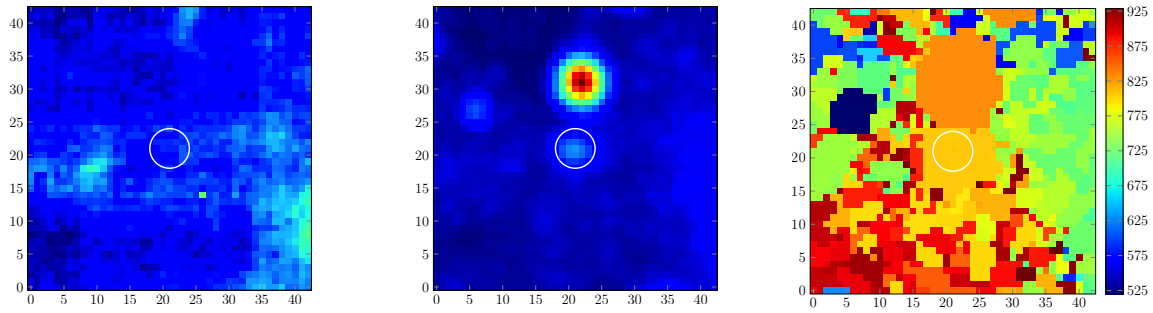
4.6.2.5 Etude de la source #243

La détection de cet objet #243 illustre la capacité de l'algorithme à détecter un signal structuré dans du bruit : bien que la raie soit localisée dans une zone du spectre très bruitée par les raies du ciel (figures 4.26a et 4.26b), la source se détache clairement du fond du ciel sur l'image des maxima après filtrage adapté et présente une homogénéité spectrale sur la carte des longueurs d'onde, figure 4.26.



(a) Spectre de l'objet #243

(b) Zoom



(c) Image blanche

(d) Valeurs maximales

(e) Longueurs d'onde

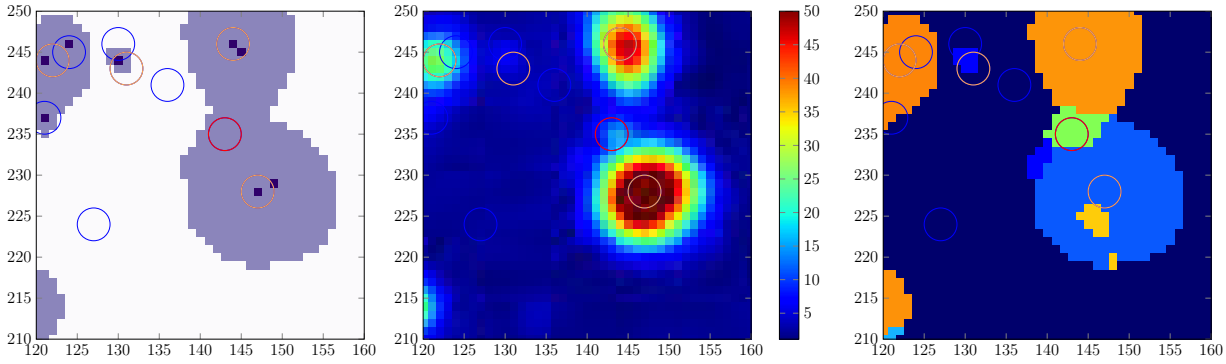
FIGURE 4.26 – Analyse de l'objet #243 avec : le spectre estimé correspondant à l'objet #243 (a) et le zoom sur la raie d'émission détectée à l'aide du max-test à la longueur d'onde $\lambda = 801.125$nm (b), le cercle blanc symbolise la position de l'objet #243 sur l'image blanche (c), la carte des valeurs maximale (d) et la carte des longueurs d'onde (e).

4.7 Améliorer la détection

L'application de l'algorithme de détection ainsi que les différents prétraitements sur le cube de données réelles HDFS a permis d'évaluer leurs performances. Si nous sommes capables de détecter des galaxies qui n'étaient pas visibles sur l'image HST, un certain nombre de fausses détections se glissent encore dans la configuration d'objets estimée. Quelques sources connues ne sont pas détectées par l'algorithme alors qu'elles sont parfaitement identifiées dans les catalogues, notamment le catalogue MUSE. Nous avons listé dans le paragraphe 4.6.1 quelques causes de non détections. Certaines ne pourront pas être résolues (détection de sources dont les centres sont superposés par exemple), mais nous pouvons travailler encore à l'amélioration de la carte de proposition, qui est responsable de la non détection de certaines sources du catalogue MUSE.

4.7.1 Perspective d'amélioration de la carte de proposition obtenue par le max-test

Certaines sources à raies d'émission du catalogue MUSE ne sont pas détectées car elles sont situées dans la périphérie d'une autre source très brillante, à spectre continu, et l'intensité de leur raie d'émission n'est pas suffisamment élevée pour que la galaxie forme un maximum local (spatial) sur la carte des maxima, même si sur la carte des longueurs d'onde la source se distingue (voir par exemple sur la figure 4.27). C'est le cas de certaines sources modélisées par des cercles rouges sur la figure 4.19 ; les pixels appartenant à ces sources sont bien classés dans $\mathcal{C}_1$ mais ne forment pas de maximum local (spatial) dans l'ensemble de pixels auquel ils appartiennent. En revanche, puisqu'une telle galaxie apparait sur la carte des longueurs d'onde comme un ensemble cohérent spatialement étendu (voir par exemple l'ensemble de pixels verts sur l'image 4.27c), nous pouvons supposer que sur l'image bande étroite centrée sur la longueur d'onde du maximum du spectre correspondant, la galaxie présente un maximum local d'intensité en son centre.



(a) Carte de proposition (b) Carte des maxima (c) Carte des longueurs d'onde

FIGURE 4.27 – Exemple d'une source, symbolisée par un cercle rouge, qui n'est pas détectable comme un maximum local, ni sur l'image blanche (a), ni sur la carte des maxima (b), mais qui apparaît comme un ensemble cohérent de pixels sur la carte des longueurs d'onde (pixels verts au centre).

Deux stratégies ont été explorées pour améliorer la détectabilité de ce genre de galaxies qui ne présentent pas de maximum local d'intensité sur la carte des maxima.

La première consiste à estimer la composante continue de chaque spectre en utilisant un filtre médian de taille 51 échantillons, de façon à obtenir un filtre 3 fois plus large que les raies d'émission de type Lyα, puis soustraire cette composante continue avant de calculer la carte des maxima et des longueurs d'onde. Ceci devrait permettre d'atténuer l'influence des galaxies brillantes à spectre continus (qui sont de toutes façons détectable sur l'image blanche

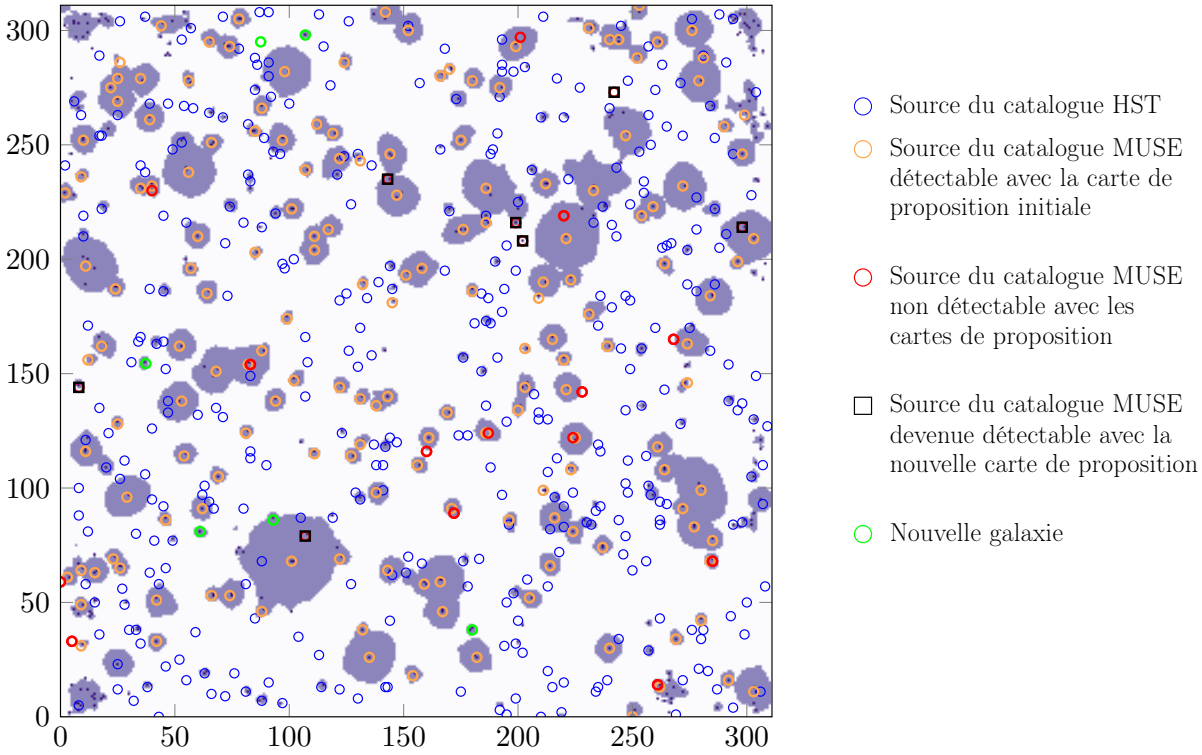
au préalable) et favoriser la recherche des maxima locaux en intensité des petites galaxies moins brillantes qui sont situées dans leur périphérie.

La deuxième stratégie utilise le seuillage de la carte des maxima obtenue à l'aide du max-test, et le même seuillage appliqué à la carte des longueurs d'onde. Les différentes longueurs d'onde se cette dernière carte sont ensuite relevées. Notons N_λ le nombre de longueurs d'onde différentes. Nous construisons ensuite N_λ tranches du cube constitués de 41 images centrées autour des N_λ longueurs d'onde relevées. Nous calculons ensuite une carte des maxima pour chacune des N_λ tranches du cube. La recherche des maxima locaux en intensité est réalisées sur les N_λ cartes des maxima. Les positions relevées sur ces N_λ cartes sont ensuite réunies pour former la carte de proposition globale.

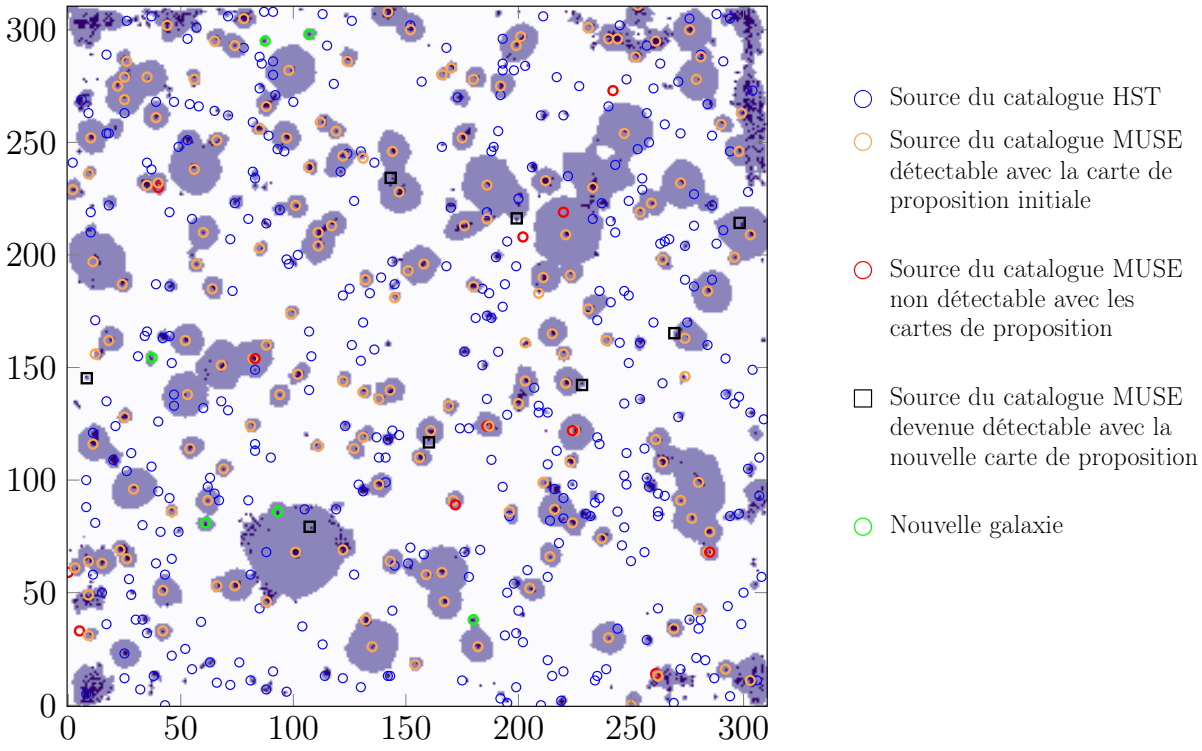
Ces deux stratégies ont été implémentées et les cartes de proposition obtenues sont présentées sur la figure 4.28.

La figure 4.28a présente la carte de détection obtenue à l'aide de la première stratégie. La composante continue de chaque spectre a été estimée puis soustraite avant le calcul de la carte des maxima. Nous constatons qu'un certain nombre de sources effectivement localisées dans l'étendue spatiale de sources à spectre continus deviennent ainsi détectables (carrés noirs), le maximum de leur profil d'intensité spatial devenant plus brillant que la composante continue de la galaxie voisine. En revanche, la soustraction de la composante continue estimée à l'aide d'un filtrage médian a entraîné la disparition de petits ensembles de pixels qui auparavant étaient supérieurs au seuil de détection, et qui voient leur intensité diminuer du fait de la soustraction. C'est notamment le cas de l'objet #184 qui, d'après l'étude menée dans le paragraphe 4.6.2.3, se révélait être effectivement une galaxie. Sur cette nouvelle carte de proposition, cet objet n'est plus proposable. En revanche le nombre de pixels candidats n'augmentant pas de façon importante, les performances en terme de fausses découvertes obtenue à l'aide de l'algorithme de détection devraient rester assez similaires à celles obtenues avec la carte de proposition déduite du max-test.

La figure 4.28b présente la carte de détection obtenue à l'aide de la seconde stratégie. Une carte des maxima à été calculée pour chacune des $N_\lambda = 994$ longueurs d'onde présentes sur la carte des longueurs d'onde seuillée (voir figure 4.9). Il faut noter qu'un certain nombre de ces longueurs d'onde correspondent aux mêmes objets (translation du maximum du spectre d'une ou deux longueurs d'onde en fonction du niveau de bruit et de l'intensité de la raie d'émission), il y aura donc une redondance parmi les N_λ cartes de proposition. Ce n'est pas vraiment problématique dans le sens où tous les pixels candidats sont ensuite réunis sur une seule et même carte. Cette méthode permet de proposer un plus grand nombre d'objets et permettraient d'améliorer le taux de détection des sources répertoriées dans le catalogue MUSE. Elle permet également de détecter séparément deux galaxies dont le centre est localisé sur le même pixel mais dont la raie d'émission est à une longueur d'onde différentes ; cependant, les limitations mathématiques (conditionnement de la matrice de configuration des galaxies $\mathbf{X}$ dans le modèle bayésien) empêcheront toujours la détection séparée des galaxies superposées. Si cette méthode permet d'améliorer le taux de détection, elle risque en revanche de détériorer le taux de fausses découvertes. Elle est en effet, beaucoup plus sensible aux structures de bruit et aux zones où la variance du bruit est plus élevée, et un grand nombre de maxima locaux vont être détectés dans ces zones à différentes positions spatiales selon les tranches de cube considérées. Ce problème est visible sur la carte 4.28b sur les bords du cube. Si en revanche, la détection est effectuée sur une troncature plus sévère du cube afin d'éviter les zones où la variance est plus élevée, cette méthode devrait fournir de très bons résultats en terme de taux de détection et de taux de fausses découvertes.



(a) Max-test appliqué aux spectres après soustraction de la composante continue.



(b) Max-test appliqué aux spectres par stack de 40 longueurs d'onde.

FIGURE 4.28 – Représentation des sources cataloguées à partir de l'image HDFS observée par le télescope spatial Hubble et des sources détectées qui n'appartiennent pas au catalogue. L'image de fond est la combinaison de la carte de proposition utilisée pour l'image blanche et la carte de proposition utilisée pour détecter les sources à raie d'émission seule.

4.7.2 Contrôler le FDR sur une liste de maxima locaux en trois dimensions

Afin d'obtenir un contrôle du FDR par objet, et non par pixel, il faudrait pouvoir définir la notion de p-valeur pour un objet détecté dans le catalogue final, ce qui constitue un problème compliqué. En admettant que les sources que nous cherchons à détecter ne possèdent qu'un maximum local dans leur réponse 3D (spatiale et spectrale), contrôler le FDR sur une liste de maxima local peut être une solution alternative. Dans les récents travaux de Schwartzman et al. [2011] et Cheng and Schwartzman [2014], les auteurs proposent de détecter des signaux rares et de faible intensité, respectivement dans des données à une dimension et à deux dimensions, à l'aide d'un algorithme de détection de maxima locaux. Ils montrent que si le bruit est stationnaire et ergodique, alors leur procédure de détection, appelée STEM pour *Smoothing and TEsting of Maxima*, basée sur la méthode de seuillage de Benjamini-Hochberg, fournit un contrôle asymptotique du FDR. La méthode STEM se décompose en quatre étape :

1. Filtrage des données pour augmenter le rapport signal à bruit des signaux d'intérêt.
2. Détection des maxima locaux supérieurs à un certain seuil de détection.
3. Calcul les p-valeurs des maxima locaux.
4. Application de la procédure de seuillage de Benjamini-Hochberg.

L'adaptation des travaux de Cheng and Schwartzman [2014] au problème de contrôle du FDR dans le catalogue d'objets estimé par notre algorithme pourrait mener à l'élaboration d'une carte de proposition dans laquelle le taux de faux maxima locaux serait garanti, et par extension, le taux de fausses découvertes dans le catalogue d'objets renvoyé par l'algorithme. Ces travaux sont laissés en perspective.

4.8 Bilan

Le modèle et les données réelles

Avec le passage aux données réelles, plusieurs écarts sont constatés entre le modèle utilisé pour définir la méthode de détection et la réalité :

- le bruit de chaque image (à λ fixée) n'est pas stationnaire, la présence de structures verticales et horizontales va à l'encontre de l'hypothèse d'un bruit gaussien i.i.d.,
- le bruit est localement corrélé spectralement et spatialement par les opérations de drizzling lors de la constitution du cube de données final.

Il a fallu développer des méthodes permettant d'approcher la loi du max-test malgré la distribution réelle des pixels de bruit dans le cube HDFS.

Résultats de la détection sur le cube HDFS

Concernant les performances de détection sur le cube HDFS, le bilan est plutôt satisfaisant puisque plus de 90% des sources répertoriées dans le catalogue MUSE ont été détectées et ce pour un nombre de fausses détections limité.
La détection effectuée à l'aide de la carte de proposition obtenue par le max-test pour une probabilité de fausse alarme de 0.1% a permis de détecter 5 nouvelles galaxies qui n'étaient pas visibles sur les données de Hubble.
Les contraintes imposées par la modélisation de la configuration d'objets (carte de proposition et d'intensité du processus ponctuel, pénalisation des recouvrements) expliquent une partie des sources non détectées.
Les améliorations de la carte de proposition laissées en perspective devraient permettre d'augmenter le taux de détection et probablement conduire à l'ajout d'autres nouvelles galaxies.

Conclusion et Perspectives

L'objectif de ce manuscrit était de présenter la méthode de détection de sources élaborée durant les trois années de thèse. Nous avons choisi d'ancrer la description de la méthode dans le contexte de la détection de galaxies lointaines dans les données hyperspectrales produites par l'instrument MUSE pour plusieurs raisons :

- les premiers travaux sur cette méthode de détection (Chatelain et al. [2011]) ont été lancés suite à un appel du consortium MUSE pour la création des nouvelles méthodes de détection de sources adaptées aux données en trois dimensions,
- cette méthode a été implémentée dans l'objectif d'être intégrée dans la suite logicielle associée aux données MUSE. Nous avons été amenés à faire des choix de modélisation en lien avec le type d'objets à détecter (étoiles et galaxies) et avec les contraintes physiques et instrumentales liées à l'application,
- cela nous a permis d'illustrer les points forts et les limitations de la méthode avec des exemples concrets qui facilitent la compréhension.

Il faut cependant garder à l'esprit que sous réserve de modifications des descripteurs des objets (forme elliptique, profil d'intensité spatial Sersic, etc), cette méthode peut s'appliquer à d'autres problématiques de détection de sources quasi-ponctuelles dans des données en trois dimensions. La troisième dimension peut-être une profondeur spatiale ou une série temporelle. Éventuellement, les objets peuvent disparaître et réapparaitre selon cet axe. Nous avons vu que cette méthode peut également s'appliquer à une image, mais de nouvelles optimisations du code, adaptées aux données en deux dimensions, devraient être implémentées afin de réduire les temps de calcul.

L'originalité de la méthode repose sur la **représentation d'une source comme un objet décrit par un faible nombre de paramètres** plutôt que comme une collection de pixels. La distribution spatiale de ces objets dans la scène observée est guidée par les prétraitements statistiques visant à extraire des données les zones de recherche, à l'aide de stratégies de type tests multiples.

L'application de l'ensemble de la chaine de traitement (normalisation, filtrage adapté, détection sur l'image blanche, détection sur le cube complet) aux données réelles HDFS a donné des résultats tout à fait satisfaisant compte-tenu des faibles rapports signal-à-bruit. Le taux de détection des sources du catalogue MUSE est supérieur à 90% pour une proportion de fausses découvertes d'environ 10%. L'algorithme a permis d'ajouter 5 nouvelles galaxies au catalogue HST. D'autres sources candidates ont pu être détectées en abaissant le seuil de construction de la carte de proposition à l'aide du max-test), mais cela au détriment de la proportion de fausses détections.

Principales contributions

Prise en compte des caractéristiques de l'instrument

L'instrument est caractérisé par sa réponse impulsionnelle, dans le cas de MUSE, il s'agit d'une réponse complexe en trois dimensions, séparable spectralement et spatialement. Devant la dimension du problème de modélisation des données et d'estimation de la configuration d'objet, nous avons été obligés de poser des hypothèses de travail simplificatrices afin de pouvoir écrire le modèle d'observation et la densité des différents paramètres du modèle. Dans le modèle d'observation, la composante spatiale (FSF) de la réponse de l'instrument est supposée identique pour toutes les longueurs d'onde, ce qui nous a permis de prendre en compte la convolution spatiale du profil d'intensité des sources par une FSF moyenne. La composante spectrale n'apparaît pas explicitement dans le modèle, elle est inclue dans l'estimation des spectres au sens du maximum *a posteriori*. La déconvolution peut être appliquée après cette étape d'estimation selon les besoins de l'utilisateur.

Si la PSF n'est pas intégrée dans sa globalité dans le processus de détection, **la variabilité spectrale de la FSF est prise en compte lors de l'étape de filtrage adapté**. Par hypothèses, les galaxies lointaines sont considérées comme des sources quasi-ponctuelles dans les trois dimensions du cube et présentent un très faible RSB. Afin d'augmenter leur RSB, il semble crucial d'inclure dans la réponse du filtre adapté à la PSF de l'instrument le plus d'informations disponibles. Nous avons vu dans le chapitre 3 que l'introduction de **ce filtrage adapté améliore la capacité de détection des sources les plus faibles**, mais annule l'hypothèse d'indépendance des pixels de bruit (dans les trois dimensions). Cette dépendance, problématique pour la plupart des méthodes de contrôle des erreurs par tests multiples, a été étudiée et prise en compte dans la conception des tests utilisés pour construire la carte de proposition des objets, qui est utilisée lors de l'échantillonnage de la configuration d'objets.

Régularisation de la configuration d'objets par prétraitements statistiques

Si l'absence de régularisation sur les intensités des objets à permis de tenir compte des grandes dynamiques entre les sources les plus brillantes et les sources les plus faibles, cela entraîne nécessairement une augmentation des fausses détections, notamment sur les données réelles où le bruit n'est pas parfaitement homogène et contient des structures d'intensité relativement élevée à certaines longueurs d'onde. Afin de limiter ces fausses détections, nous avons mis en place **une phase de prétraitement** qui a pour but, d'une part, de **limiter la proposition des objets aux zones les plus susceptibles de contenir la contribution d'une source**, et d'autre part, de **contrôler un taux d'erreur global sur cette étape de pré-détection**. Deux types de contrôles ont été étudiés :

— le contrôle du FWER sur chaque spectre à travers le max-test qui consiste à décider si un spectre appartient à une source ou non en fonction de la valeur maximale du spectre après filtrage adapté. Cela revient à contrôler une probabilité de fausse alarme globale pour la statistique de la valeur maximum du spectre,

— le contrôle du FDR sur le cube entier à l'aide de la procédure de seuillage de Benjamini-Hochberg en utilisant les conditions de dépendance énoncées par Benjamini and Yekutieli [2001]. Nous avons montré que le contrôle du FDR pouvait être appliqué au résultat d'un outil fréquemment utilisé en traitement du signal, le filtrage adapté, sous quelques conditions très simples (positivité de la réponse du filtre et positivité des sources) permettant d'utiliser les résultats de Benjamini and Yekutieli [2001].

Implémentation de la méthode et transfert au consortium MUSE

Une autre part importante du travail de thèse, qui n'apparaît pas dans ce manuscrit, est l'**implémentation et la validation de la méthode de détection et de tous les traitements proposés en langage Python**. Le code a été conçu pour fonctionner avec les packages Python dédiés à la manipulation des données MUSE, mais sera également distribué de manière indépendante à la communauté traitement du signal. A l'heure actuelle, le code a été transféré au consortium MUSE, et intégré à la suite logicielle *mpdaf* dédiées aux données MUSE sous le nom de SELFI pour *Source with Emission Line FInder*. **L'avantage de ce code est sa modularité**, puisqu'il suffit de coder une nouvelle classe objet (où sont définies les caractéristiques géométriques, spectrales, d'intensité des objets du processus ponctuels marqués) pour être utilisé dans un autre contexte applicatif. Le modèle d'observation est également géré par une classe indépendante qui peut être modifiée selon les besoins. Il faut noter que si le code peut fonctionner aussi bien sur des images (deux dimensions) que des données en trois dimensions, il a été spécialement optimisé pour ce dernier cas. Les traitements réalisés longueurs d'onde par longueurs d'onde sont effectués en parallèle, ce qui nécessite une architecture du code particulière, qui entraîne des lenteurs lors de l'utilisation sur une seule image.

Perspectives

Prise en compte d'*a priori* sur la présence de sources connues pour affiner la détection

Une des limitations de l'algorithme est la détection des sources brillantes, spatialement étendue et dont le support est trop complexe pour être modélisé par une ellipse avec une décroissance d'intensité centre-bord de type Sersic. Il en est de même pour les étoiles localisées dans le champ d'observation de l'instrument, il s'agit de sources parfaitement ponctuelles, leur réponse doit donc être identique à la réponse impulsionnelle de l'instrument, qui dans le cas de MUSE est modélisée par une fonction Moffat en deux dimensions sur un support circulaire. Les contraintes de forme et de profil d'intensité spatial n'étant pas les mêmes que pour les autres sources, la modélisation par un profil Sersic à support elliptique ne convient pas. De plus l'étoile la plus brillante du champ (ou les étoiles lorsqu'il y en a plusieurs) est utilisée pour estimer la PSF, la position et les caractéristiques de cette étoile sont donc parfaitement connues. Nous avons donc initialisé l'algorithme de détection avec ces informations afin d'éviter les problèmes de détections multiples sur le support de l'étoile. Lorsque le champ observé contient des galaxies proches connues, nous pourrions étendre l'initialisation à ces sources. **Les informations de modélisation pourraient être extraites d'études complémentaires** (observations avec d'autres instruments, estimation des caractéristiques à partir de l'image blanche avec des méthodes développées pour les images simples). **Les résidus de modélisation au niveau de ces sources brillantes seraient alors largement atténués** et cela permettrait de **réduire le nombre de fausses détections**.

L'amélioration de la carte de proposition

L'influence de la carte de proposition ne doit pas être négligée puisqu'elle est utilisée comme un seuillage dur des données. Si certaines galaxies n'ont pas été détectées lors de la phase de prétraitement par tests multiples, la carte de proposition ne contiendra pas de pixels au niveau de leur localisation et aucun objet ne pourra être proposé lors de l'échantillonnage RJMCMC de la configuration d'objet pour détecter ces galaxies. A la fin du chapitre 4, **nous avons proposé quelques améliorations de la carte de proposition** construite à l'aide du max-test. Il faudrait maintenant qualifier la détection résultant de ces nouvelles cartes en terme de taux de

détection et de taux de fausses découvertes. Les approches par tests multiples nécessitant une modélisation des données sous l'hypothèse de bruit seul, la non-homogénéité des pixels de bruit sur les données MUSE entraîne des erreurs de détection et dégrade potentiellement la carte de proposition. Nous pouvons espérer que les prochaines versions du cube HDFS, ainsi que les futurs observations contiendront moins d'artefacts dus à l'instrument, ou à la chaîne de réduction de données et que la carte de proposition sera donc plus précise.

Contrôle d'un taux d'erreur dans le catalogue d'objets

En partant du principe que les données contiennent peu d'artefacts et que l'algorithme accepte la naissance d'un objet au niveau de chaque pixel de la carte de détection, sauf lorsque les centres sont trop proches l'un de l'autre et ne respectent pas le critère de Rayleigh (qui est utilisé comme pénalisation hard core dans la densité du processus ponctuels), **si l'on parvient à contrôler un critère de type FDR sur ces pixels candidats, nous aurons également un contrôle de type FDR pour la liste des objets détectés**. Pour cela, il faudrait par exemple apprendre la loi des maxima locaux sous l'hypothèse nulle. Afin d'apprendre la loi des maxima locaux, nous pourrions utiliser la loi des minima locaux, qui sous hypothèses de symétrie de la distribution du bruit et de positivité des sources, devrait être identique (au signe près) à la loi des maxima sous l'hypothèse nulle. Le problème provient surtout du fait que sur une carte des minima locaux de la taille des données MUSE (environ 300×300 pixels), le nombre de minima locaux n'est pas suffisamment important pour apprendre précisément la loi et donc en déduire les p-valeurs correspondantes et transposer l'analyse à la carte des maxima locaux. Une autre solution serait d'étendre les travaux de Schwartzman et al. [2011] (1D) et Cheng and Schwartzman [2014] (2D) aux données en trois dimensions, car les objets qui nous intéressent, constituent un maximum local en intensité (leur position spatiale dans le champ) et spectralement (position de la raie d'émission). Là encore, une solution serait d'utiliser les minima locaux pour déterminer (au signe près) la loi des maxima locaux puisque les pixels de bruit sont corrélés et sont ensuite filtrés lors de l'étape de filtrage adapté.

Annexes

Table des matières

Annexe A

Principales définitions liées aux tests d'hypothèses

En statistiques, un test d'hypothèse est un problème d'inférence statistique consistant à valider ou à rejeter une hypothèse statistique à partir de l'observation x (scalaire ou vectorielle) d'un phénomène physique. La procédure de rejet d'une hypothèse nécessite la mise en place d'un test, *i.e.* une fonction T de l'observation x, dont la valeur sera ensuite comparée à un seuil de décision permettant de rejeter ou de ne pas rejeter l'hypothèse considérée. Dans cette annexe, nous considérerons tout d'abord le cas d'un test d'hypothèses binaire, qui consiste à choisir entre deux hypothèses concurrentes, $\mathcal{H}_0$ et $\mathcal{H}_1$, pour expliquer l'observation x. Dans la seconde partie de cette annexe, nous considérerons le cas des tests multiples, qui consiste à mener l'analyse d'acceptation-rejet des hypothèses pour un ensemble de N observations $\boldsymbol{x} = [x_1, \cdots, x_N]$.

A.1 Une introduction au test d'hypothèses binaire

La construction d'un test d'hypothèses binaire débute par la définition des deux hypothèses, $\mathcal{H}_0$ et $\mathcal{H}_1$, de la statistique du test et du seuil permettant de décider si le résultat du test est significatif pour l'une ou l'autre des hypothèses.

A.1.1 Mise en place d'un test d'hypothèses binaire

Par convention, pour la détection d'une source, l'hypothèse $\mathcal{H}_0$, appelée hypothèse nulle, modélise le cas où l'observation x ne contient que du bruit, et l'hypothèse $\mathcal{H}_1$, appelée hypothèse alternative, est le complément de l'hypothèse nulle :

$$\begin{cases} \mathcal{H}_0 & : \quad x \sim p(x|\mathcal{H}_0) \quad \text{(bruit seul)} \\ \mathcal{H}_1 & : \quad x \sim p(x|\mathcal{H}_1) \quad \text{(source + bruit)} \end{cases}, \tag{A.1}$$

où $p(x|\mathcal{H}_i)$ est la distribution de l'observation x sous l'hypothèse $\mathcal{H}_i$. Une fois les hypothèses définies, il faut mettre en place une méthode de décision pour valider l'une ou l'autre des hypothèses. Pour cela, il faut définir une statistique de test $T(\cdot)$ et un seuil de décision associé à cette statistique. La statistique de test $T(\cdot)$ doit expliquer au mieux les données x afin de permettre à l'utilisateur de de rejeter ou ne pas rejeter l'hypothèse nulle. Afin d'être exploitable, cette statistique de test doit avoir une densité de probabilité analytiquement définie sous l'hypothèse $\mathcal{H}_0$ et éventuellement sous $\mathcal{H}_1$. La règle de décision $\mathcal{D}$ associée au test $T(\cdot)$ et à la valeur de seuil η se résume par :

$$T(x) \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\lessgtr}} \eta, \tag{A.2}$$

Le choix du seuil η permet de contrôler la probabilité de faire une erreur de décision pour un test $T(\cdot)$ donné.

A.1.2 Contrôler le résultat d'un test d'hypothèses binaire

En appliquant la règle de décision $\mathcal{D}$ associée à un test $T(\cdot)$ quatre situations sont à envisager, elles sont décrites dans le tableau A.1. Connaissant la statistique du test $T(\cdot)$, une probabilité peut être associée à chacune des décisions.

Décision / Vérité	$\mathcal{H}_0$ est retenue	$\mathcal{H}_0$ est rejetée
$\mathcal{H}_0$ est vraie	✓ $1 - p_{FA}$	Erreur de type 1 Fausse alarme p_{FA}
$\mathcal{H}_0$ est fausse	Erreur de type 2 Détection manquée p_M	✓ Puissance du test $p_D = 1 - p_M$

TABLEAU A.1 – Probabilités associées aux différentes décisions possibles.

La probabilité de fausses alarmes p_{FA} caractérise la probabilité de rejeter l'hypothèse $\mathcal{H}_0$ sachant qu'elle est vraie. Pour notre problème de détection, elle s'exprime alors sous la forme :

$$p_{FA} = \Pr(T(x) > \eta | \mathcal{H}_0) = \int_{\eta}^{+\infty} p(T(x)|\mathcal{H}_0) dx. \tag{A.3}$$

Si la distribution du test sous $\mathcal{H}_1$ est connue, il est possible de déterminer également la probabilité de détection manquée p_M :

$$p_M = \Pr(T(x) < \eta | \mathcal{H}_1) = \int_{-\infty}^{\eta} p(T(x)|\mathcal{H}_1) dx. \tag{A.4}$$

A.1.2.1 Exemple

Supposons que x est une réalisation d'une variable aléatoire X gaussienne. On suppose que le signal d'intérêt est porté par une moyenne strictement positive pour cette variable aléatoire. Déterminer si le signal est présent revient à tester ces deux hypothèses :

$$\begin{cases} \mathcal{H}_0 & : & x \sim \mathcal{N}(0, \sigma^2) & & \text{(bruit seul)} \\ \mathcal{H}_1 & : & x \sim \mathcal{N}(\theta, \sigma^2) \ , & \theta > 0 & \text{(source + bruit)} \end{cases} . \tag{A.5}$$

Le test le plus simple dans ce cas est $T(x) = x$, dans ce cas la statistique du test sous chacune des hypothèses est la statistique de l'hypothèse elle-même. La figure A.1 illustre la statistique du test sous les deux hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$ ainsi que la représentation des probabilités de fausses alarmes et de détections manquées correspondant à un seuil de décision η. La fonction de répartition de la loi normale $\mathcal{N}(0, 1)$ est connue et se calcule numériquement :

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{t^2}{2}\right) dt.$$

Sous l'hypothèse $\mathcal{H}_0$, le test $T(x) = x$ suit une loi gaussienne $\mathcal{N}(0, \sigma^2)$ de variance, sa fonction de répartition s'écrit dont :

$$F_{\mathcal{H}_0}(x) = \Phi\left(\frac{x}{\sigma}\right)$$

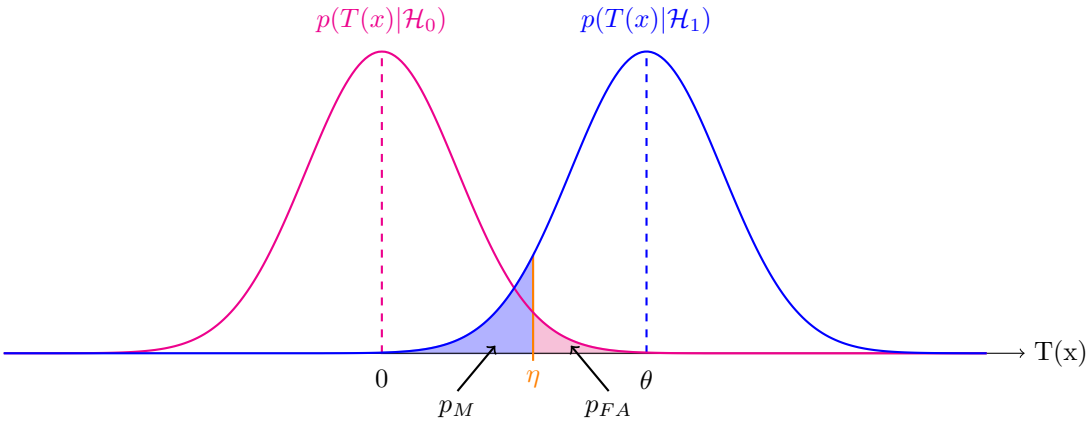


FIGURE A.1 – Représentation des statistiques du test $T(x) = x$ sous les deux hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$

Pour une probabilité de fausse alarme $p_{FA} = \alpha$ fixée, le seuil η se déduit alors simplement de α comme :

$$\frac{\eta}{\sigma} = F^{-1}_{\mathcal{H}_0}(1-\alpha)$$

De même, la probabilité de détection manquée p_M s'écrit en fonction du seuil de décision η de la façon suivante :

$$p_M = F_{\mathcal{H}_1}(\eta) = \Phi\left(\frac{x-\theta}{\sigma}\right) \tag{A.6}$$

A.1.2.2 Contrôle des erreurs de type 1

Plaçons-nous dans le cas d'un test d'hypothèses binaire $T(\cdot)$ pour lequel nous souhaitons contrôler les erreurs de type 1. Le test consiste alors à tester uniquement la validité de l'hypothèse nulle pour l'observation x. Pour notre problème de détection, rejeter l'hypothèse $\mathcal{H}_0$ avec un niveau de confiance $1-\alpha$ revient à déterminer le seuil η tel que :

$$p_{FA} = \int_{\eta}^{+\infty} p(T(x)|\mathcal{H}_0)dx \leqslant \alpha,$$

or

$$p_{FA} = 1 - \int_{-\infty}^{\eta} p(T(x)|\mathcal{H}_0)dx = 1 - F_{\mathcal{H}_0}(\eta),$$

où $F_{\mathcal{H}_0}$ est la fonction de répartition du test $T(\cdot)$ sous $\mathcal{H}_0$. Trouver le seuil η tel que la probabilité de fausses alarmes soit inférieure ou égale à α revient à résoudre :

$$F_{\mathcal{H}_0}(\eta) = 1-\alpha$$

Si la fonction de répartition $F_{\mathcal{H}_0}$ du test sous $\mathcal{H}_0$ n'est pas analytiquement connue, elle peut être approchée par méthode de Monte Carlo en simulant un grand nombre de données sous l'hypothèse nulle et en calculant le test pour chacun des échantillons. La fonction de répartition empirique sera alors utilisée à la place de $F_{\mathcal{H}_0}$.

Il faut également noter que ce type de contrôle ne nous donne pas d'information sur la puissance du test, *i.e.* sur la probabilité de détection associée à ce test, si la distribution du test

sous l'hypothèse alternative n'est pas connue. Il sera en revanche la seule méthode de contrôle lorsque le test consiste à accepter ou rejeter l'hypothèse nulle sans connaissance *a priori* sur la forme du signal x sous l'hypothèse alternative.

Par convention, on rejette toujours l'hypothèse $\mathcal{H}_0$. Pour contrôler les erreurs de type 2, il suffit d'inverser les rôles de $\mathcal{H}_0$ et $\mathcal{H}_1$. Notons qu'il n'est pas possible de contrôler simultanément les erreurs de type 1 et les erreurs de type 2.

A.1.3 Définition de la p-valeur

La p-valeur est un outil fréquemment utilisé en analyse statistique. La p-valeur associée à une observation x_i désigne la probabilité pour $T(x)$ d'être au moins aussi extrême que le résultat $T(x_i)$ du test sur l'observation x_i si l'hypothèse $\mathcal{H}_0$ est vraie. Pour notre problème de détection, la p-valeur s'écrit :

$$p_{x_i} = \Pr(T(x) > T(x_i)|\mathcal{H}_0) = \int_{T(x_i)}^{+\infty} p(T(x)|\mathcal{H}_0)dx. \tag{A.7}$$

On peut aussi exprimer la p-valeur de la façon suivante :

$$p_{x_i} = 1 - P(T(x) < T(x_i)|\mathcal{H}_0).$$

En notant $F_{\mathcal{H}_0}$ la fonction de répartition de la statistique de test sous $\mathcal{H}_0$, l'équation précédente devient :

$$p_{x_i} = 1 - F_{\mathcal{H}_0}(T(x_i)). \tag{A.8}$$

La p-valeur est une probabilité, c'est donc une variable appartenant à l'intervalle $[0, 1]$. Les p-valeurs sont obtenues comme une transformation de la statistique de test. Elles permettent d'obtenir un outil universel tel que :

- p_{x_i} est distribuée selon une loi uniforme $\mathcal{U}([0, 1])$ sous $\mathcal{H}_0$,
- p_{x_i} est stochastiquement plus petite que $\mathcal{U}([0, 1])$ si $x_i \sim \mathcal{H}_1$, *i.e.* $Pr(p_{x_i} < t) > t$ pour tout $t \in [0, 1]$,

et ceci indépendamment de la loi initiale de $T(x_i)$.

A.1.3.1 Exemple

Pour le test d'hypothèses binaire décrit par le modèle (A.5), la figure A.2 illustre graphiquement le calcul de la p-valeur associé à l'observation x_i.
La p-valeur est déterminée par l'équation :

$$p_{x_i} = 1 - \Phi\left(\frac{x_i}{\sigma}\right).$$

A.2 Une introduction aux tests multiples

Supposons maintenant que nous disposons d'un ensemble de N observations $\boldsymbol{x} = [x_1, \cdots, x_N]$. A chaque observation x_i est associé un système d'hypothèses binaire :

$$\begin{cases} \mathcal{H}_0^i & : \quad x \sim p(x_i|\mathcal{H}_0^i) \\ \mathcal{H}_1^i & : \quad x \sim p(x_i|\mathcal{H}_1^i) \end{cases}$$

Nous allons considérer ici le cas où les hypothèses $\mathcal{H}_0^i$ (resp. $\mathcal{H}_1^i$) sont identiques pour les N observations, donc $\mathcal{H}_0^i = \mathcal{H}_0$ (resp. $\mathcal{H}_1^i = \mathcal{H}_1$). Les résultats suivants restent cependant valides pour des hypothèses $\mathcal{H}_0^i$ et $\mathcal{H}_1^i$ différentes selon les observations x_i considérées. Le tableau A.2 représente la répartition des décisions prises pour les N observations à l'aide des N tests correspondants. Le

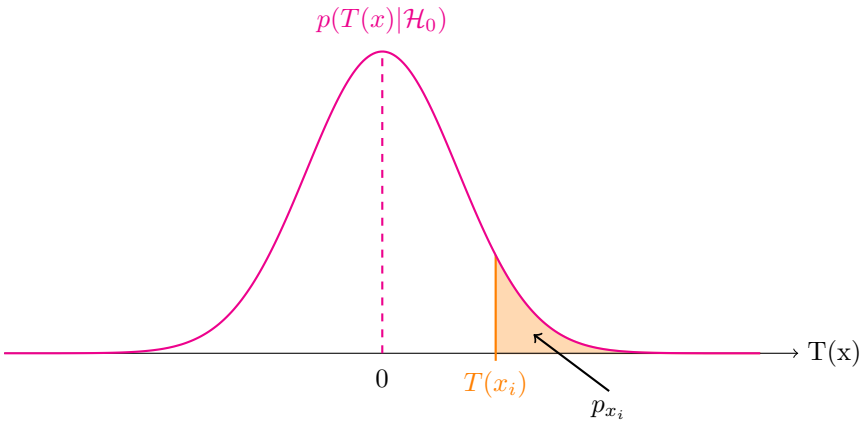


FIGURE A.2 – Représentation graphique du calcul de la p-valeur associée au test $T(x_i)$ effectué sur l'observation gaussienne x_i.

Décision / Vérité	$\mathcal{H}_0$ est retenue	$\mathcal{H}_0$ est rejetée	Total
$\mathcal{H}_0$ est vraie	N_0 - a	a	N_0
$\mathcal{H}_0$ est fausse	N_1 - b	b	N_1
Total	N_0 + N_1 - R	R	N

TABLEAU A.2 – Répartition des N décisions associées aux N tests.

nombre N_0 (resp. N_1) représente le nombre de cas qui sont modélisés par l'hypothèse $\mathcal{H}_0$ (resp. $\mathcal{H}_1$). Le taux de fausses alarmes sur l'ensemble des N décisions est représenté par la fraction a/N et le nombre $R = a + b$ représente le nombre de cas qui ont été classés sous l'hypothèse $\mathcal{H}_1$.

La problématique dans le cas de ces N tests peut se formuler de façon différente que dans le cas d'un unique test. Bien sûr, il est possible de contrôler chaque test de manière individuelle en contrôlant la probabilité de fausses alarmes ou de détections manquées. Il peut cependant être bien plus intéressant d'avoir un critère de contrôle global des erreurs sur l'ensemble des N tests. Pour bien comprendre la problématique, considérons le cas où la probabilité de fausses alarmes est contrôlée de manière individuelle pour les N tests à un niveau α. Le nombre moyen de fausses alarmes sur l'ensemble des tests sera alors de $N\alpha$. Si le nombre de tests est très grand alors le nombre de fausses alarmes sera également important, indépendamment du nombre b de cas correctement détectés sous l'hypothèse $\mathcal{H}_1$. Dans la littérature deux grands critères de contrôle global des erreurs dans le cas de tests multiples ont été introduits : le contrôle du FWER pour *family wise error rate* et le contrôle du FDR pour *false discovery rate*.

A.2.1 Contrôle du FWER

Le critère du FWER consiste à contrôler la probabilité d'effectuer au moins une fausse découverte parmi les N tests à un niveau α :

$$FWER = \Pr(a \geqslant 1) \leqslant \alpha \tag{A.9}$$

Cette probabilité peut s'interpréter comme une probabilité de fausse alarme globale définie pour l'ensemble des tests. On trouve dans la littérature différentes procédures de contrôle du FWER, nous avons choisi d'en présenter deux, la correction de Bonferroni et la méthode de Holm-Bonferroni introduites lors des premières tentatives de contrôle global des erreurs dans le cas de tests multiples.

A.2.1.1 Correction de Bonferroni

La correction de Bonferroni énonce le principe suivant : rejeter tous les cas où la p-valeur $p_{x_i} < \frac{\alpha}{N}$ permet de maintenir le $FWER \leqslant \alpha$. Ceci est démontré à l'aide de l'inégalité de Boole :

$$FWER = P\left(\bigcup_{i \in I_0} \left(p_{x_i} \leqslant \frac{\alpha}{N}\right)\right) \leqslant \sum_{i \in I_0} P\left(p_{x_i} \leqslant \frac{\alpha}{N}\right), \tag{A.10}$$

où l'ensemble I_0 représente l'ensemble des cas où $\mathcal{H}_0$ est vraie. Si $\mathcal{H}_{0,i}$ est vraie alors la p-valeur associée p_{x_i} suit une loi uniforme sur l'intervalle $[0, 1]$ et donc $P\left(p_{x_i} \leqslant \frac{\alpha}{N}\right) = \frac{\alpha}{N}$. Finalement :

$$FWER \leqslant \sum_{i \in I_0} \frac{\alpha}{N} = \frac{N_0 \alpha}{N} \leqslant \alpha \tag{A.11}$$

Cette procédure est bien trop conservative puisque le seuil dépend du nombre N de tests considérés.

A.2.1.2 Méthode séquentielle de descente de Holm-Bonferroni

Holm [1979] propose de raffiner ce résultat à l'aide de la procédure de Holm-Bonferroni décrite dans l'encadré A.1 :

Encadré A.1 – Méthode séquentielle de descente de Holm-Bonferroni

1. Soit $\mathcal{H}_0^{(1)}$, ..., $\mathcal{H}_0^{(N)}$ une famille d'hypothèses nulles et $p_1, ..., p_N$ les p-valeurs correspondantes.
2. Notons $p_{(1)} \leqslant ... \leqslant p_{(N)}$ ces p-valeurs ordonnées de façon croissante et $\mathcal{H}_0^{(1)}$, ..., $\mathcal{H}_0^{(N)}$ les hypothèses nulles associées.
3. Pour un niveau d'importance α, soit i_{min} l'indice minimal tel que :
$$i_{min} = \underset{i}{\operatorname{argmin}}\ p_{(i)} \geqslant \frac{\alpha}{N + 1 - i}$$
4. Rejet des hypothèses nulles $\mathcal{H}_0^{(1)}$, ..., $\mathcal{H}_0^{(i_{min}-1)}$ et acceptation des hypothèses $\mathcal{H}_0^{(i_{min})}$, ..., $\mathcal{H}_0^{(N)}$.

La procédure décrite dans l'encadré A.1 permet finalement de contrôler le FWER à un niveau α : $FWER \leqslant \alpha$. La méthode de Holm-Bonferroni est uniformément plus puissante que la méthode de Bonferroni du fait du seuil adaptatif $\frac{\alpha}{N+1-i}$ qui augmente à chaque nouvelle p-valeur testée alors que le seuil reste constant pour la méthode de Bonferroni.

A.2.2 Contrôle du FDR

Le contrôle du FDR a été introduit par Benjamini and Hochberg [1995]. La proportion de fausses découvertes correspondant aux N tests dont la répartition est décrite dans le tableau A.2 s'écrit :

$$\text{FDP} = \frac{a}{R}, \tag{A.12}$$

en posant $FDP = 0$ si R vaut 0, i.e. si toutes les hypothèses nulles sont vraies. Le taux de fausses découverte, FDR, est donné par :

$$\text{FDR} = \text{E}\Big[\text{FDP}\Big] = \text{E}\Big[\frac{a}{R}\Big]. \tag{A.13}$$

Contrôler le taux de fausses découvertes, c'est-à-dire maintenir en moyenne la FDP en dessous d'un seuil q signifie assurer que sur les R hypothèses nulles rejetées, la proportion d'hypothèses rejetées à tort est en moyenne inférieure à q. La procédure de Benjamini and Hochberg [1995] permet de contrôler le FDR dans le cas de N tests indépendants à un niveau $\pi_0 q$ où $\pi_0 = \frac{N_0}{N}$ est la proportion de tests réellement sous l'hypothèse nulle, et $0 \leqslant q \leqslant 1$ est le paramètre de contrôle. Si la proportion $0 \leqslant \pi_0 \leqslant 1$ n'est pas toujours connue, le contrôle est toujours garanti à un niveau q. La procédure de Benjamini-Hochberg est détaillée dans l'encadré A.2.

ENCADRÉ A.2 – Procédure de Benjamini-Hochberg

1. Soit $\mathcal{H}_0^{(1)}$, ..., $\mathcal{H}_0^{(N)}$ une famille d'hypothèses nulles et $p_1, ..., p_N$ les p-valeurs correspondantes.
2. Par convention, on pose $p_{(0)} = 0$ et l'on note $p_{(0)} < p_{(1)} \leqslant ... \leqslant p_{(N)}$ les p-valeurs ordonnées de façon croissante et $\mathcal{H}_0^{(1)}$, ..., $\mathcal{H}_0^{(N)}$ les hypothèses nulles associées.
3. Soit $k = \underset{i}{\text{argmax}} \left(p_{(i)} \leqslant q\frac{i}{N}\right)$
4. Rejet des hypothèses nulles $\mathcal{H}_0^{(1)}$, ..., $\mathcal{H}_0^{(k)}$ et acceptation des hypothèses $\mathcal{H}_0^{(k+1)}$, ..., $\mathcal{H}_0^{(N)}$.

Puisque le FDR est un critère moins conservatif que le FWER, la détection effectuée avec une procédure de contrôle du FDR sera donc plus puissante. A noter que toute procédure qui contrôle le FWER contrôle également le FDR, mais de façon sous-optimale.

A.2.2.1 Exemple

Soit $N = 500$ échantillons simulés indépendamment selon une loi gaussienne $\mathcal{N}(\theta, 1)$ où $\theta = 0$ sous $\mathcal{H}_0$ et $\theta > 0$ sous $\mathcal{H}_1$. A chaque échantillon le modèle d'hypothèses binaire suivant est associé :

$$\begin{cases} \mathcal{H}_0^i & : \quad x_i \sim \mathcal{N}(0, \sigma^2) & \text{(bruit seul)} \\ \mathcal{H}_1^i & : \quad x_i \sim \mathcal{N}(\theta, \sigma^2)\ , \quad \theta > 0 & \text{(source + bruit)} \end{cases}.$$

La proportion d'échantillons simulés suivant $\mathcal{H}_0$ est $\pi_0 = 0.9$ ($N_0 = 450$). Les p-valeurs correspondantes sont calculées comme dans l'exemple développé dans le paragraphe A.1.3 et la procédure de Benjamini-Hochberg est appliquée pour détecter les hypothèses $\mathcal{H}_1$ tout en limitant les fausses découvertes. La figure A.3 illustre le principe de la procédure sur les N échantillons considérés ici, pour un contrôle au niveau $\pi_0 q$ avec $q = 0.2$.

Pour un contrôle du FWER sur les mêmes échantillons x_i avec la procédure de Holm-Bonferroni, la détection des échantillons générés sous $\mathcal{H}_1$ est beaucoup moins puissante qu'avec la

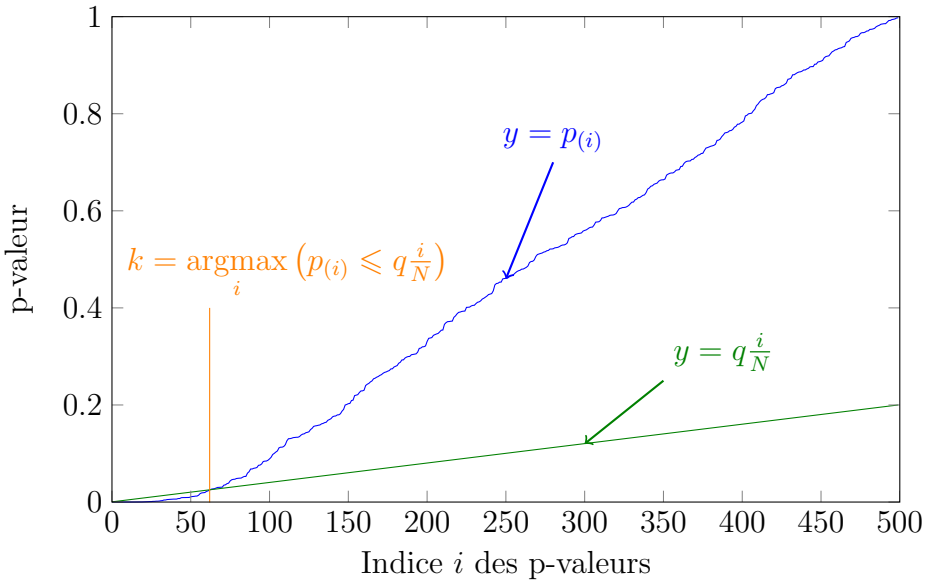


FIGURE A.3 – Représentation graphique du seuillage des p-valeurs avec la procédure de Benjamini-Hochberg sur $N = 500$ échantillons gaussiens indépendants. Dans cet exemple $k = 62$ pour un niveau de contrôle $q = 0.2$.

procédure de Benjamini-Hochberg. Si aucune hypothèse $\mathcal{H}_0$ n'a été rejetée à tort pour différents niveaux de contrôle α du FWER ($\alpha = 20\%$, $\alpha = 10\%$ et $\alpha = 5\%$), en revanche seulement 30% à 40% des hypothèses $\mathcal{H}_1$ ont été correctement détectées. La figure A.4 illustre le seuillage des p-valeurs à l'aide de la procédure de Holm-Bonferroni.

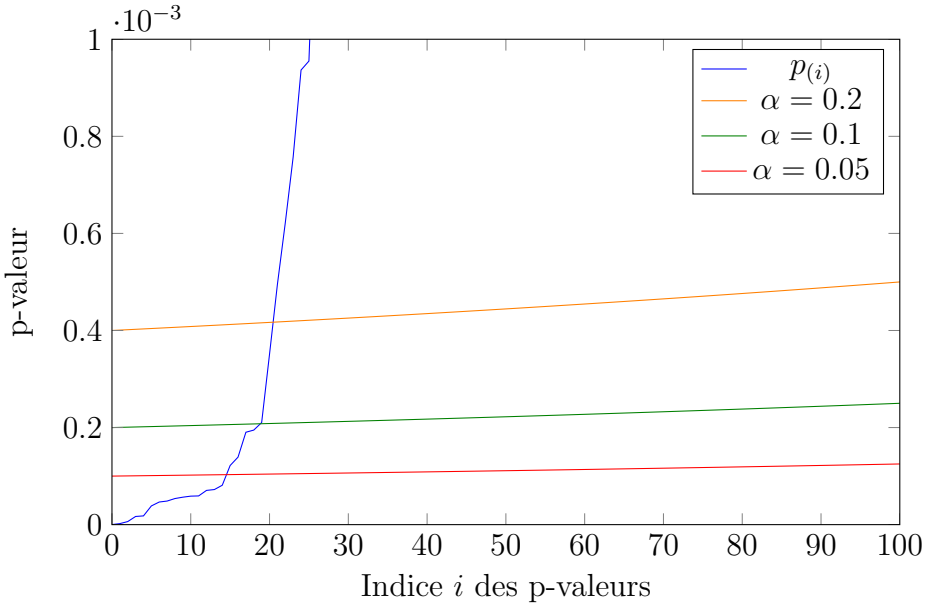


FIGURE A.4 – Représentation graphique du seuillage des p-valeurs avec la procédure de Holm-Bonferroni sur $N = 500$ échantillons gaussiens indépendants. Seules les premières p-valeurs ordonnées sont représentées ainsi que les seuillages pour différents niveaux de contrôle α.

Annexe B

Les processus ponctuels marqués et leur utilisation en imagerie

Cette annexe aurait pû être un chapitre à part entière du manuscrit. L'étude théorique des processus ponctuels marqués n'est pas fondamentale à la compréhension du fonctionnement de la méthode de détection des galaxies dans les données MUSE. Nous avons donc fait le choix de placer ce chapitre en annexe, à l'intention du lecteur intéressé par la théorie des processus ponctuels marqués et à leur simulation.

Cette annexe se décompose en trois parties. Les principales définitions et propriétés des processus ponctuels sont définies dans la première partie. Une attention particulière sera portée au processus de Poisson qui sert de processus de référence à d'autres processus plus complexe. Ce processus de Poisson sera notamment utilisé lors de la construction des processus ponctuels marqués qui permettent de modéliser les configurations d'objets que nous cherchons à détecter. La deuxième partie est consacrée à la simulation des processus ponctuels marqués, différents algorithmes permettant d'échantillonner ces processus y sont détaillés. Enfin dans une dernière partie, nous donnerons un aperçu des différentes étapes et précautions à prendre pour la modélisation de la configuration d'objets recherchée à l'aide d'un processus ponctuel marqué.

B.1 Processus ponctuel marqués

Nous allons introduire dans cette partie les principales définitions et propriétés des processus ponctuels marqués qui seront utilisées pour traiter le problème de détection de galaxies dans les données MUSE. Une description détaillée des processus ponctuels marqués peut se trouver dans différents ouvrages tels que Van Lieshout [2000] ou Chiu et al. [2013].

B.1.1 Espace des configurations

Soit $\mathcal{P}$ un ensemble non vide muni d'une distance d tel que $(\mathcal{P}, d)$ soit complet et séparable. On appelle point tout élément $u \in \mathcal{P}$.

Définition B.1. *On appelle configuration* $\boldsymbol{u}$ *un ensemble dénombrable de points non ordonné de* $\mathcal{P}$ *:*

$$\boldsymbol{u} = \{u_1, u_2, \cdots, u_n\}, n \in \mathbb{N}$$

Dans le cadre de la modélisation de configurations d'objets dans les images, on considérera uniquement les processus définis sur des régions bornées : le support de l'image (composée de pixels) projetée dans un espace continu, avec un nombre fini de points. On définit une configuration localement finie et simple de la manière suivante : $\boldsymbol{u} \in \mathcal{P}$ est localement finie si dans tout

borélien borné $A \subset \mathcal{P}$, elle place un nombre fini de points distincts $N_{\boldsymbol{u}}(A) = n$. L'ensemble des configurations de points localement finies et simples (les points sont tous distincts) est noté N^{lf}.

Définition B.2. *L'ensemble de toutes les configurations localement finies et simples dans $\mathcal{P}$ est noté :*

$$\Omega = \bigcup_{n \in \mathbb{N}} \Omega_n \tag{B.1}$$

où $\Omega_n = \left\{\{u_1, ..., u_n\}; u_i \in \mathcal{P} \quad \forall 1 \leq i \leq n\right\}$ est l'ensemble des configurations de n points non ordonnés.

Pour les applications en imagerie traditionnelle (deux dimensions), $\mathcal{P} = \mathbb{R}^2$ et d la distance euclidienne associée à $\mathcal{P}$, dans ce cas, une configuration est une distribution spatiale aléatoire de points.

B.1.2 Processus ponctuels : définitions et notations

Définition B.3. *Un processus ponctuel sur $\mathcal{P}$ est donc une application U de l'espace probabilisé dans N^{lf} telle que pour tout borélien $A \in \mathcal{P}$, $N_{\boldsymbol{u}}(A)$ est une variable aléatoire presque surement finie.*

On définit l'intensité d'un processus ponctuel comme le moment d'ordre un de la variable aléatoire $N_{\boldsymbol{u}}(A)$. La mesure $\nu(\cdot)$ sur $\mathcal{P}$, appelée **mesure d'intensité** est :

$$\nu(A) = \mathbb{E}\left[N_{\boldsymbol{u}}(A)\right], \text{ pour tout borélien } A \in \mathcal{P} \tag{B.2}$$

On parle de processus ponctuel stationnaire si sa mesure d'intensité est invariante par translation i.e.

$$\nu(A + v) = \mathbb{E}\left[N_{\boldsymbol{u}}(A + v)\right] = \mathbb{E}\left[N_{\boldsymbol{u}}(A)\right] = \nu(A), \text{ pour tout } v \in \mathcal{P} \tag{B.3}$$

Or les seules mesures invariantes par translation sont les multiples de la mesure de Lebesgue, ce qui donne lieu au théorème suivant : si ν est invariante par translation sur $\mathbb{R}^d$ alors : $\nu(A) = c \times \Lambda(A)$, $\forall c > 0$, où $\Lambda(\cdot)$ est la mesure de Lebesgue. La mesure de Lebesgue est une mesure qui étend le concept de volume à tout produit cartésien d'intervalles bornés $I_1 \times I_2 \times \cdots \times I_n \subseteq \mathbb{R}^n$.

B.1.3 Processus de Poisson

Soit $\boldsymbol{u}$ une réalisation aléatoire d'un processus ponctuel dont les points sont distribués de façon indépendante.

Définition B.4. *On dit qu'un processus ponctuel est un processus de Poisson si pour tout borélien borné A de $\mathcal{P}$:*

- *$N_{\boldsymbol{u}}(A)$ est le nombre de points de $\boldsymbol{u}$ qui tombent dans A et $N_{\boldsymbol{u}}(A)$ est une variable aléatoire qui suit une loi de Poisson discrète d'espérance $\nu(A)$,*
- *pour k boréliens disjoints $A_1, \ldots, A_k$, les variables $N_{\boldsymbol{u}}(A_1), ..., N_{\boldsymbol{u}}(A_k)$ sont indépendantes.*

La figure B.1 montre un exemple de réalisation d'un processus ponctuel de Poisson homogène sur une image, modélisée comme une région bornée de $\mathbb{R}^2$. Le processus de Poisson homogène est un cas particulier des processus de Poisson pour lequel la mesure d'intensité $\nu(\cdot)$ est proportionnelle à la mesure de Lebesgue : $\nu(\cdot) = \lambda\Lambda(\cdot)$, d'intensité constante $\lambda \in \mathbb{R}^{+*}$. Pour tout borélien borné $A \subset \mathcal{P}$, il suffit alors de tirer le nombre de points selon une loi de Poisson d'espérance $\nu(\mathcal{P}) = \lambda\Lambda(\mathcal{P})$ et de les répartir de manière uniforme sur A (dans l'exemple de la figure B.1, A représente l'image de taille $P \times Q$ pixels).
Lorsque l'on possède des informations *a priori* sur la localisation des objets que l'on cherche à

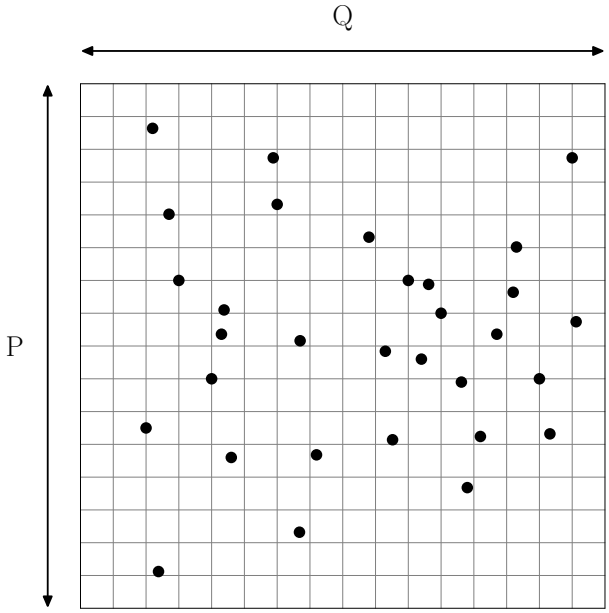


FIGURE B.1 – Exemple de réalisation d'un processus ponctuel de Poisson homogène sur une région bornée de $\mathbb{R}^2$.

modéliser, il peut être intéressant d'injecter cette information directement dans la mesure d'intensité du processus ponctuel. Les processus ponctuels de Poisson non homogènes permettent de considérer une répartition non uniforme des points sur $\mathcal{P}$. La mesure d'intensité $\nu(\cdot)$ n'est plus uniforme sur $\mathcal{P}$. Elle est obtenue en considérant, non plus une intensité constante λ, mais une fonction d'intensité non négative $\lambda(\cdot)$ définie comme la dérivée de Radon-Nikodym de $\nu(\cdot)$ par rapport à la mesure de Lebesgue $\Lambda(\cdot)$:

$$\forall A \in \mathcal{P}, \quad \nu(A) = \int_A \lambda(\boldsymbol{u})\Lambda(d\boldsymbol{u}) < \infty \tag{B.4}$$

Il est possible d'injecter dans cette fonction $\lambda(\cdot)$ des informations sur la localisation des points du processus ponctuel, extraites directement des observations. Dans ce cas, la mesure d'intensité $\nu(\cdot)$ du processus de référence permet de favoriser les zones où l'intensité est plus élevée. La figure B.2 montre une réalisation d'un processus de Poisson non homogène dont la fonction d'intensité est modélisée par les zones de couleur. L'intensité de ce processus est cinq fois plus élevée dans la zone rose que dans la zone orange.

Quel que soit le type de processus de Poisson considéré (homogène ou non homogène) on peut maintenant définir sa mesure de probabilité π_ν. Pour tout borélien $B \subset \Omega$ la mesure de probabilité associée au processus de Poisson est :

$$\pi_\nu(B) = e^{-\nu(\mathcal{P})} \sum_{n=0}^{+\infty} \frac{\pi_{\nu_n}(B)}{n!} \tag{B.5}$$

avec

$$\pi_{\nu_n}(B) = \begin{cases} \mathbb{1}_{[\emptyset \in B]}, \text{ si } n = 0, \\ \int_{\mathcal{P}} \dots \int_{\mathcal{P}} \mathbb{1}_{[\{u_1,\dots,u_n\} \in B_n]} \nu(du_1)\dots\nu(du_n), \quad \text{pour } n \geq 1, \end{cases} \tag{B.6}$$

où $\mathbb{1}_{[A]}$ est la fonction indicatrice pour A (1 si A est vrai, 0 sinon), $\nu(\cdot)$ est la mesure d'intensité du processus et B_n est un sous-ensemble de configurations dans B avec exactement n points.

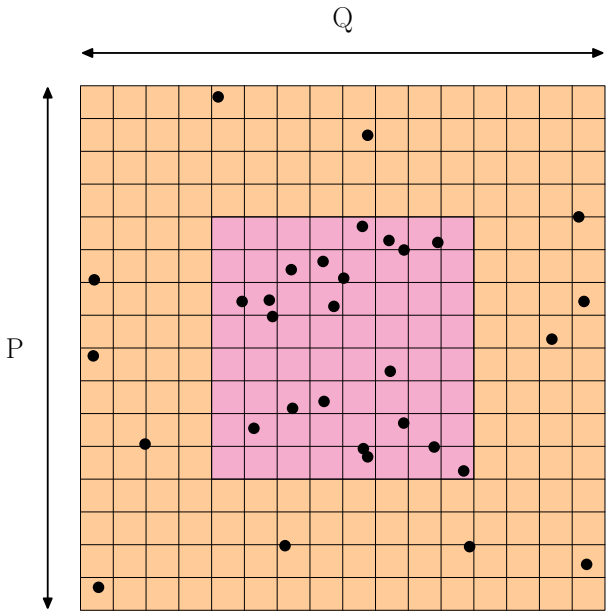


FIGURE B.2 – Exemple de réalisation d'un processus ponctuel de Poisson non homogène sur une région bornée de $\mathbb{R}^2$. Cette configuration place plus de points au centre de l'image (zone rose), là où l'intensité est plus élevée que sur les bords (zone orange).

Finalement, le processus ponctuel de Poisson peut être paramétré par l'intensité $\beta = \nu(\mathcal{P})$ et par la mesure d'intensité normalisée $\nu'(A) = \nu(A)/\nu(\mathcal{P})$ sur $\mathcal{P}$. On a alors :

$$\pi_{\nu'_n}(B) = \frac{\pi_{\nu_n}(B)}{\beta^n}, \quad \text{pour } n \geq 0, \tag{B.7}$$

ce qui conduit, pour tout borélien $B \subset \Omega$ à la mesure suivante :

$$\pi_\nu(B) \equiv \pi_\beta(B) = e^{-\beta} \sum_{n=0}^{+\infty} \frac{\beta^n \pi_{\nu'_n}(B)}{n!}, \tag{B.8}$$

où le paramètre d'intensité β représente le nombre moyen de points des réalisations du processus de Poisson sur $\mathcal{P}$.

B.1.4 Densité d'un processus ponctuel

Les processus de Poisson permettent de construire une grande famille de processus ponctuels. On définit la densité de ces processus en fonction de la mesure de probabilité $\pi_\nu(\cdot)$ d'un processus ponctuel de Poisson de référence. Une telle densité est une fonction non négative $f(\cdot)$ définie sur l'espace des configurations Ω telle que :

$$1 = \int_\Omega f(\boldsymbol{u}) d\pi_\nu(\boldsymbol{u}). \tag{B.9}$$

L'équation (B.9) montre que la densité $f(\cdot)$ du processus s'exprime comme la dérivée de Radon-Nikodym de la mesure de probabilité par rapport à la mesure du processus de Poisson de référence π_ν. La mesure de probabilité du processus s'écrit sur chaque borélien $B \subset \Omega$: $P(B) = \int_B f(\boldsymbol{u}) d\pi_\nu(\boldsymbol{u})$. Par exemple, la mesure de probabilité $\pi_\beta(B)$ du processus ponctuel de Poisson

de paramètre d'intensité β donné dans l'équation (B.8) s'écrit :

$$\begin{aligned}\pi_\beta(B) &= e^{-\beta}\sum_{n=0}^{+\infty}\frac{\beta^n}{n!}\int_B d\pi_{\nu'_n}(\boldsymbol{u}),\\ &= \int_B e^{-\beta}\beta^{n(\boldsymbol{u})}\sum_{n=0}^{+\infty}\frac{d\pi_{\nu'_n}(\boldsymbol{u})}{n!},\\ &= \int_B e^{1-\beta}\beta^{n(\boldsymbol{u})}\,d\pi_{\nu'}(\boldsymbol{u}),\end{aligned}$$

où la dernière ligne est obtenue à partir de l'équation (B.5) puisque par construction $\nu'(\mathcal{P}) = 1$. Par identification avec l'équation (B.9), le processus admet une densité

$$f(\boldsymbol{u}|\beta) = \beta^{n(\boldsymbol{u})}\exp(1-\beta),\ \forall \boldsymbol{u} \in \Omega, \tag{B.10}$$

définie par rapport au processus de Poisson normalisé $\pi_{\nu'}$ (mesure qui intègre l'information d'intensité $\lambda(\cdot)$ définie dans l'équation (B.4)).

Notons qu'en général il n'est pas possible d'exprimer analytiquement l'intégrale sur Ω dans l'équation (B.9). Par conséquent, ces processus sont définis par une densité non normalisée $p(\cdot)$ telle que la densité normalisée s'écrive $f(\cdot) = p(\cdot)/c$, avec $c = \int_\Omega p(\boldsymbol{u})d\pi_\nu(\boldsymbol{u})$.

B.1.5 Processus ponctuels marqués

Dans le cadre de la détection d'objets dans une image, il est intéressant de modéliser la position des objets dans l'image par un processus ponctuel auquel on associe un ensemble de marques caractéristiques de la géométrie, de l'intensité ou de toutes autres caractéristiques de l'objet d'intérêt.

Définition B.5. *Un processus ponctuel marqué sur $\mathcal{X} = \mathcal{P} \times \mathcal{M}$ est un processus ponctuel sur l'espace des positions $\mathcal{P} = \mathbb{R}^d$ auquel sont ajoutées des marques. C'est une variable aléatoire dont les réalisations sont des configurations aléatoires d'objets dans l'ensemble $\mathcal{X}$. Le processus ponctuel sur $\mathcal{P}$ possède les mêmes propriétés que celles décrites dans la section B.1.2.*

Une configuration d'objets $\boldsymbol{u}$ est donc définie de la façon suivante :

$$\boldsymbol{u}_\mathcal{M} = \{(u_1, m_1), (u_2, m_2), \cdots, (u_n, m_n)\}, n \in \mathbb{N}$$

où m_i désigne l'ensemble des marques associées au point u_i. Dans la suite de ce manuscrit, nous ne considérerons plus que des processus ponctuels marqués, afin d'alléger les notations, nous noterons $\boldsymbol{u}$ la configuration d'objet et u_i un objet de cette configuration (u_i = position + marques). Par analogie avec les processus ponctuels, un processus ponctuel marqué peut être représenté par sa densité de probabilité non normalisée $h(\cdot)$ que nous détaillerons dans la partie B.3.

B.2 Simulation des processus ponctuels marqués

Dans le cadre bayésien choisi (chapitre 2), l'extraction de configuration d'objets, modélisées par processus ponctuels marqués, nécessite de simuler ces processus selon le modèle de densité choisi, puis d'extraire une configuration d'objets à partir d'un estimateur adapté. Différents types d'algorithmes d'échantillonnage des processus ponctuels marqués sont proposés dans la littérature : les algorithmes de naissance/mort de Baddeley and Van Lieshout [1993], les algorithmes de simulation exacte de Kendall and Møller [2000], les processus à temps continus de Descombes

et al. [2009], les algorithmes d'échantillonnage à sauts réversibles auxquels nous allons nous intéresser dans cette partie. La méthode de détection de galaxies que nous proposons dans le chapitre 2 repose sur la modélisation de la configuration de galaxies par un processus ponctuel marqué qui doit être échantillonné. Le nombre de galaxies est *a priori* inconnu, le processus doit donc être échantillonné dans un espace de dimension variable, ce que permettent les algorithmes d'échantillonnages à sauts réversibles. Les algorithmes d'échantillonnage à sauts réversibles présentent l'avantage de travailler en dimension variable, de pouvoir mélanger différents types de mouvements (ajout, suppression, modification d'objets, modification de paramètres du modèle).

B.2.1 Rappel des notations et remarques générales sur les échantillonneurs proposés

Dans les paragraphes suivants nous allons décrire différents types d'échantillonneurs. Afin de mieux visualiser les différences entre ces algorithmes dans le cadre de l'échantillonnage d'un processus ponctuel marqué, nous allons utiliser les notations introduites précédemment. Nous rappelons les principales notations liées aux processus ponctuels marqués :

- soit $p(\cdot)$ la densité non normalisée du processus ponctuel marqué définie par rapport à $\pi_\nu(\cdot)$,
- $\pi_\nu(\cdot)$ représente la mesure de probabilité associée au processus,
- $\Omega = \bigcup_{n\in\mathbb{N}} \Omega_n$ est l'espace des configurations,
- $\mathcal{P}$ est un ensemble non vide qui permet par exemple de décrire le support d'une image,
- $\boldsymbol{u}$ est une configuration d'objets telle que $\boldsymbol{u} = \big\{\{u_1, ..., u_n\}; u_i \in \mathcal{P}\big\}$,
- dans les algorithmes d'échantillonnage itératifs, à l'itération k, $\boldsymbol{u} = \boldsymbol{u}^{(k-1)}$ désignera la configuration courante et $\boldsymbol{v}$ la configuration proposée lors de cette itération. Une étape d'acceptation-rejet permettra de décider si $\boldsymbol{u}^{(k)} = \boldsymbol{u}$ ou si $\boldsymbol{u}^{(k)} = \boldsymbol{v}$.

Une fois le processus ponctuel marqué défini par sa mesure de probabilité $\pi_\nu(B)$ pour tout borélien $B \subset \Omega$, il faut l'échantillonner suivant sa densité $p(\cdot)$. Nous avons soulevé dans le paragraphe B.1.4 que bien souvent, la densité du processus ponctuel marqué est définie à une constante de normalisation près. Il n'est donc pas possible de simuler directement une réalisation de ce processus. Les méthodes d'échantillonnage basées sur des algorithmes de type MCMC permettent de simuler un processus stochastique à partir de sa densité non normalisée ; elles permettent de construire une chaîne de Markov qui converge vers la loi $p(\cdot)$ du processus. Les algorithmes d'échantillonnage que nous allons présenter dans les paragraphes suivants ont été proposés dans le cadre très général de l'échantillonnage d'une loi de probabilité.

Historiquement, les travaux de Metropolis et al. [1953] et Hastings [1970] ont mené à la construction de l'algorithme d'échantillonnage de Metropolis-Hastings qui permet la simulation d'un processus à partir de sa densité non normalisée dans un espace de dimension fixe, ce qui suppose que le nombre d'objets de la configuration à simuler est connu. Geyer and Møller [1994] ont ensuite proposé un algorithme de naissance/mort basé sur le principe de l'échantillonneur de Metropolis-Hastings qui permet d'échantillonner le processus en dimension variable, en autorisant l'ajout et la suppression d'objets. Enfin Green [1995] a étendu le principe de l'échantillonnage en dimension variable introduit par Geyer and Møller [1994] à d'autres mouvements que les naissances et les morts. Ces différents algorithmes sont présentés dans les paragraphes suivants, et nous introduirons un échantillonneur de Metropolis-Hastings particulier : l'échantillonneur de Gibbs, qui est utilisé dans la méthode de détection de galaxies décrite au chapitre 2 pour échantillonner les paramètres du modèle des données (moyenne et variance du bruit) qui apparaissent dans la fonction de vraisemblance.

B.2.2 Algorithme de type Metropolis-Hastings

L'algorithme de Metropolis-Hastings permet l'échantillonnage d'un ensemble de paramètres $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)$ dans un espace de dimension n fixe. L'idée directrice de l'algorithme MCMC de Metropolis-Hastings tel qu'il a été proposé dans les travaux de Metropolis et al. [1953] et Hastings [1970] repose sur la construction d'une chaine de Markov $(\boldsymbol{\theta}^{(k)})_{k\in\mathbb{N}}$ dont la distribution stationnaire est la densité de probabilité $p(\boldsymbol{\theta})$ du vecteur de paramètres $\boldsymbol{\theta}$, définie à une constante de normalisation près. La convergence en loi de la chaîne de Markov vers la loi cible $p(\boldsymbol{\theta})$ est assurée par une étape d'acceptation-rejet de la perturbation proposée.

Dans le cadre de la simulation d'une réalisation $\boldsymbol{u}$ d'un processus ponctuel marqué, l'échantillonnage de Metropolis-Hastings a lieu dans un sous espace de l'ensemble Ω_n des configurations localement finie de dimension n, où n correspond au nombre d'objets de la configuration à simuler. La loi cible est la densité non normalisée $p(\cdot)$ du processus, la chaîne de Markov $(\boldsymbol{u}^{(k)})_{k\in\mathbb{N}}$ générée doit converger vers cette loi et permettre ainsi l'estimation de la configuration d'objets qui maximise cette loi. A chaque itération k de l'algorithme, une nouvelle configuration $\boldsymbol{v}$, issue de la perturbation de la configuration courante $\boldsymbol{u} = \boldsymbol{u}^{(k-1)}$, est proposée à l'aide d'une loi de proposition $q(\boldsymbol{v}, \boldsymbol{u}) = q(\boldsymbol{v}|\boldsymbol{u})$ (ou loi instrumentale) que nous devons savoir simuler. Cette loi de proposition peut représenter par exemple une marche aléatoire sur l'un des paramètres décrivant la configuration $\boldsymbol{u}$ (la position ou l'une des marques d'un objet $u_i \in \boldsymbol{u}$). Le principe de l'échantillonneur de Metropolis-Hastings est présenté dans l'encadré B.1.

ENCADRÉ B.1 – Echantillonneur de Metropolis-Hastings

1. **Initialisation :** $\boldsymbol{u}^{(0)}$ est une configuration d'objets arbitraire.
2. **Iteration** k **:**
   — Générer $\boldsymbol{v} \sim q(\boldsymbol{v}|\boldsymbol{u})$
   — Calcul du ratio :
   $$r(\boldsymbol{v}, \boldsymbol{u}) = \frac{p(\boldsymbol{v})q(\boldsymbol{u}|\boldsymbol{v})}{p(\boldsymbol{u})q(\boldsymbol{v}|\boldsymbol{u})} \tag{B.11}$$
3. **Etape d'acceptation-rejet :**
   $$\boldsymbol{u}^{(k)} = \begin{cases} \boldsymbol{v} \text{ avec la probabilité } \min(r(\boldsymbol{v}, \boldsymbol{u}), 1) \\ \boldsymbol{u} \text{ sinon} \end{cases} \tag{B.12}$$

Dans la pratique, lors de l'étape d'acceptation-rejet, une variable aléatoire α est générée selon une loi uniforme $\mathcal{U}_{[0,1]}$, si $\alpha \leq r(\boldsymbol{v}, \boldsymbol{u})$ alors la valeur $\boldsymbol{v}$ est acceptée sinon, la valeur $\boldsymbol{u}$ est conservée. Le ratio $r(\boldsymbol{v}, \boldsymbol{u})$ prend en compte le rapport de vraisemblance à travers le terme $\frac{p(\boldsymbol{v})}{p(\boldsymbol{u})}$. L'étape d'acceptation est aléatoire, ce qui assure la convergence de la chaîne de Markov vers la loi cible $p(\cdot)$.

B.2.3 Algorithme de naissance/mort de Geyer et Møller

Lorsque le nombre d'objets de la configuration n est un des paramètres à estimer, il est nécessaire d'explorer l'espace Ω des configurations simples et finies, défini par l'équation (B.1). Pour ce faire, Geyer and Møller [1994] ont proposé un algorithme d'échantillonnage pour les processus ponctuels spatiaux qui utilise des perturbations de type naissance-mort afin d'explorer l'espace Ω. Nous rappelons le principe de cet algorithme de type naissance/mort dans l'encadré 2.2.

Encadré B.2 – Algorithme de type naissance et mort de Geyer et Møller

Notons $\frac{\nu(\cdot)}{\nu(\mathcal{P})}$ la mesure d'intensité normalisée d'un processus ponctuel.

1. Soit $\boldsymbol{u}^{(k-1)} = \boldsymbol{u}$ la configuration courante, à l'itération k, comportant $n(\boldsymbol{u})$ points.
2. Choix d'une naissance avec une probabilité $\frac{1}{2}$ ou d'une mort avec une probabilité $\frac{1}{2}$.
3. **Naissance :** Proposition d'un objet $v \in \mathcal{P}$ généré selon $\frac{\nu(\cdot)}{\nu(\mathcal{P})}$ à ajouter à la configuration courante $\boldsymbol{u}$. Soit $\boldsymbol{v} = v \cup \boldsymbol{u}$, calculer le taux :
$$r(\boldsymbol{u}, \boldsymbol{v}) = \frac{p(\boldsymbol{v})}{p(\boldsymbol{u})} \frac{\nu(\mathcal{P})}{n(\boldsymbol{v})}$$
4. **Mort :** Choix d'un objet u à supprimer de la configuration courante $\boldsymbol{u}$. Soit $\boldsymbol{v} = \boldsymbol{u}/u$, calculer le taux :
$$r(\boldsymbol{u}, \boldsymbol{v}) = \frac{p(\boldsymbol{v})}{p(\boldsymbol{u})} \frac{n(\boldsymbol{u})}{\nu(\mathcal{P})}$$
5. Accepter la proposition $\boldsymbol{u}^{(k)} = \boldsymbol{v}$ avec la probabilité $\min(1, r(\boldsymbol{u}, \boldsymbol{v}))$, rejeter la proposition sinon : $\boldsymbol{u}^{(k)} = \boldsymbol{u}$.

La structure de l'algorithme est très similaire à celle de l'échantillonneur de Metropolis-Hastings, mais tandis que ce dernier permet de modifier les marques ou la position des objets de la configuration d'objet, l'algorithme de Geyer and Møller [1994] propose à chaque itération d'ajouter ou de retirer un objet de la configuration courante. Ainsi au lieu de modifier un objet existant pour l'ajuster au mieux aux données, la stratégie développée consiste à supprimer cet objet et à en proposer un nouveau, avec des marques et une position légérement différentes plus tard dans le processus d'échantillonnage. Cependant cette stratégie est coûteuse, en terme de temps de convergence et en terme de calcul, d'où l'utilisation d'échantillonneur de Metropolis-Hastings-Green présenté dans le paragraphe suivant.

B.2.4 Echantillonneur de Metropolis-Hastings-Green

L'échantillonnage de type Metropolis-Hastings a été généralisé à des espaces de dimension variable par Green [1995] sous le nom de méthode de Monte Carlo par chaîne de Markov à sauts réversibles, RJMCMC pour *Reversible Jump Monte Carlo Markov Chain*, également appelée échantillonneur de Metropolis-Hastings-Green. Cet échantillonneur permet notamment de mettre en oeuvre des perturbations sur la configuration d'objet plus compliquées que les naissances et les morts utilisées dans l'algorithme de Geyer and Møller [1994]. La structure de l'échantillonneur, présentée dans l'encadré B.3, assez similaire à celle de l'algorithme de Metropolis-Hastings diffèrent par les points suivants :

1. la densité non normalisée $p(\cdot)$ est remplacée dans le ratio d'acceptation-rejet par une mesure non normalisée $\pi_\nu(\cdot)$ sur l'espace $\Omega = \bigcup_{n \in \mathbb{N}} \Omega_n$,
2. la loi de proposition $q(\boldsymbol{v}, \boldsymbol{u})$ est remplacée par un noyau de proposition réversible $\mathcal{Q}(\boldsymbol{v}, d\boldsymbol{u})$ tel que :
$$\int_A \int_B \pi(d\boldsymbol{u}) \mathcal{Q}(\boldsymbol{u}, d\boldsymbol{v}) = \int_B \int_A \pi(d\boldsymbol{v}) \mathcal{Q}(\boldsymbol{v}, d\boldsymbol{u})$$

Avec ces modifications, le ratio d'acceptation-rejet devient le ratio de Metropolis-Hastings-Green défini par :

$$r(\boldsymbol{v}, \boldsymbol{u}) = \frac{\pi(d\boldsymbol{v})\mathcal{Q}(\boldsymbol{v}, d\boldsymbol{u})}{\pi(d\boldsymbol{u})\mathcal{Q}(\boldsymbol{u}, d\boldsymbol{v})}$$

ENCADRÉ B.3 – Echantillonneur de Metropolis-Hastings-Green

1. **Initialisation :** $\boldsymbol{u}^{(0)}$ est une configuration d'objets arbitraire.
2. **Iteration** k **:**
   — Générer $\boldsymbol{v} \sim \mathcal{Q}(\boldsymbol{v}, \boldsymbol{u})$
   — Calcul du ratio :
   $$r(\boldsymbol{v}, \boldsymbol{u}) = \frac{\pi(d\boldsymbol{v})\mathcal{Q}(\boldsymbol{v}, d\boldsymbol{u})}{\pi(d\boldsymbol{u})\mathcal{Q}(\boldsymbol{u}, d\boldsymbol{v})} \tag{B.13}$$
3. **Etape d'acceptation-rejet :**
   $$\boldsymbol{u}^{(k)} = \begin{cases} \boldsymbol{v} \text{ avec la probabilité } \min(r(\boldsymbol{v}, \boldsymbol{u}), 1) \\ \boldsymbol{u} \text{ sinon} \end{cases} \tag{B.14}$$

Intéressons nous au mouvement de naissance/mort et notons $\mathcal{Q}_{BD}$ [1] le noyau de naissance-mort tel que :

$$\mathcal{Q}_{BD}(\boldsymbol{u}, \cdot) = p_B(\boldsymbol{u})\mathcal{Q}_B(\boldsymbol{u}, \cdot) + p_D(\boldsymbol{u})\mathcal{Q}_D(\boldsymbol{u}, \cdot),$$

avec p_B (respectivement p_D) la probabilité de proposer une naissance (respectivement une mort) et $\mathcal{Q}_B$ (respectivement $\mathcal{Q}_D$) le sous-noyau de naissance (respectivement de mort). Le ratio de Metropolis-Hastings-Green associé au mouvement de naissance est :

$$r(\boldsymbol{v}, \boldsymbol{u}) = \frac{p_D(\boldsymbol{v})}{p_B(\boldsymbol{u})} \frac{p(\boldsymbol{v})}{p(\boldsymbol{u})} \frac{\nu(\mathcal{P})}{n(\boldsymbol{u}) + 1}$$

où la nouvelle configuration $\boldsymbol{v}$ s'écrit $\boldsymbol{v} = \boldsymbol{u} \cup v$ avec v l'objet à ajouter. Le ratio de Metropolis-Hastings-Green associé au mouvement de mort est :

$$r(\boldsymbol{v}, \boldsymbol{u}) = \frac{p_B(\boldsymbol{v})}{p_D(\boldsymbol{u})} \frac{p(\boldsymbol{v})}{p(\boldsymbol{u})} \frac{n(\boldsymbol{u})}{\nu(\mathcal{P})}$$

Si $p_B = p_D = \frac{1}{2}$ alors on retrouve bien l'algorithme de naissance/mort de Geyer and Møller [1994] décrit dans l'encadré B.2. Le détail des calculs des ratios est donné dans le livre de Descombes [2011].

Des noyaux caractérisant des mouvements plus compliqués peuvent être utilisés, on peut par exemple créer un noyau de fusion/division ou encore de naissance/mort dans un voisinage. Les mouvements de fusion/division peuvent être obtenus par combinaison de noyau de naissance et de mort, on peut interpréter la fusion comme la mort de deux objets et la naissance d'un nouvel objet créé à partir des deux objets supprimés ou encore comme la mort d'un des deux objets à fusionner et la modification de la position et des marques de l'objet restant. En pratique, dans la méthode proposée dans le chapitre 2, nous n'avons pas implémenté de mouvement de fusion/division, notamment parce que les naissances sont favorisées dans les zones les plus probables grâce à une intensité non uniforme du processus ponctuel utilisé, mais ces mouvements ont été envisagés dans des travaux antérieurs.

1. B pour *birth* et D pour *death*.

B.2.5 Echantillonneur de Gibbs

L'échantillonneur de Metropolis-Hastings présenté dans le paragraphe B.2.2, utilise une loi de proposition q pour simuler le vecteur de paramètres $\boldsymbol{\theta}$, indépendante de la loi $p(\cdot)$ du vecteur de paramètres $\boldsymbol{\theta}$. Si cette loi instrumentale est mal calibrée, la convergence vers la loi cible peut être très lente. Un échantillonneur couramment utilisé dans les méthodes de type MCMC est l'échantillonneur de Gibbs introduit par Geman and Geman [1984]. L'échantillonnage de Gibbs est un cas particulier de l'échantillonnage de Metropolis-Hastings : la loi $p(\boldsymbol{\theta})$ est directement utilisée lorsque la distribution conditionnelle de chaque paramètre est connue étant donnés les autres paramètres, ces lois conditionnelles jouent alors le rôle de la loi de proposition q dans le ratio de Metropolis-Hastings. Le mouvement proposé sera alors toujours accepté puisque le ratio vaudra toujours 1 (voir le détail des calculs dans [Descombes, 2011, chapitre 4]). Sans perte de généralité, on considérera ici un vecteur de deux paramètres $\boldsymbol{\theta} = (\theta_1, \theta_2)$ et leur loi conditionnelle respective :

$$\begin{aligned} \theta_1 &\sim p_1(\theta_1|\theta_2) \\ \theta_2 &\sim p_2(\theta_2|\theta_1) \end{aligned} \tag{B.15}$$

Si les deux lois conditionnelles p_1 et p_2 sont connues et que l'on sait simuler des variables selon ces deux lois alors il est alors possible d'utiliser l'algorithme d'échantillonnage de Gibbs présenté dans l'encadré B.4 pour échantillonner le vecteur de paramètres $\boldsymbol{\theta} = (\theta_1, \theta_2)$ selon $p(\boldsymbol{\theta})$.

ENCADRÉ B.4 – Echantillonneur de Gibbs

Initialisation : $\theta_1^{(0)}$ et $\theta_2^{(0)}$ fixés à une valeur arbitraire.

Iteration k **:** Connaissant $\theta_1^{(k-1)}$ et $\theta_2^{(k-1)}$, générer :

1. $\theta_1^{(k)}$ selon $p_1(\theta_1|\theta_2^{(k-1)})$
2. $\theta_2^{(k)}$ selon $p_2(\theta_2|\theta_1^{(k)})$

Il est possible de démontrer par récurrence que si $\boldsymbol{\theta}^{(k-1)} = (\theta_1^{(k-1)}, \theta_2^{(k-1)})$ est distribué selon la loi jointe cible $p(\boldsymbol{\theta})$ alors $\theta_2^{(k-1)}$ est distribué selon la loi marginale $p(\theta_2)$ et $(\theta_1^{(k)}, \theta_2^{(k-1)})$ est aussi distribué selon $p(\boldsymbol{\theta})$. Le même raisonnement est appliqué à la seconde étape de l'échantillonnage et $\boldsymbol{\theta}^{(k)} = (\theta_1^{(k)}, \theta_2^{(k)})$ est distribué selon la loi jointe cible $p(\boldsymbol{\theta})$. La démonstration est donnée dans [Robert, 2006, chapitre 6].

L'avantage de ce type d'échantillonnage repose sur des simulations de processus univariés, mais impose de connaître toutes les lois conditionnelles de $p(\boldsymbol{\theta})$. Ce type d'échantillonnage est particulièrement bien adapté aux modèles hiérarchiques de Bayes. Prenons l'exemple d'un paramètre θ à simuler selon la loi a posteriori $p(\theta|y)$ avec θ qui dépend d'un hyperparamètre λ, la loi a posteriori s'écrit alors :

$$p(\theta|y) = \int_{\Lambda} p_1(\theta|y, \lambda) p_2(\lambda|y) d\lambda \tag{B.16}$$

Il faut alors échantillonner conjointement le paramètre θ et l'hyperparamètre λ selon la procédure présentée dans l'encadré B.4.

B.3 Application à l'extraction de configurations d'objets dans les images

L'application des processus ponctuels marqués pour la détection de configurations d'objets dans des images a été largement explorée dans le domaine de la télédétection, voir par exemple les travaux de Ortner [2004], Keresztes et al. [2009] ou encore Descamps et al. [2008].

Considérons ici le problème de détection d'une configuration d'objets suivant :

— le nombre d'objets est inconnu,
— la position de chaque objet est inconnue,
— la forme paramétrique de l'objet est connue (modèle simple décrit par un petit nombre de paramètres), mais les valeurs des paramètres sont inconnues,
— les relations entre les objets peuvent être modélisées par une densité qui dépend de l'application considérée.

Afin de détecter la configuration d'objets observée dans une image bruitée, nous allons commencer par modéliser ces objets à l'aide d'un processus ponctuel marqué caractérisé par une densité $p(\cdot)$ qu'il nous faudra échantillonner afin de réaliser ensuite une estimation de la configuration.

B.3.1 Modéliser une configuration d'objets par un processus ponctuel marqué

La modélisation d'une configuration d'objets par un processus ponctuel marqué dans les problèmes de détection d'objets en imagerie a été introduite dans [Baddeley and Van Lieshout, 1993, chapitre 11] et dans [Molina and Ripley, 1993, chapitre 13]. Le livre de Descombes [2011] présente le lien entre processus ponctuels marqués et application à la détection de configurations aléatoire d'objets dans des images ; le lecteur pourra se reporter à cet ouvrage pour une description complète des méthodes liées à l'utilisation de processus ponctuels marqués en imagerie.

En imagerie, les données $\mathbf{Y} \in \mathbb{R}^d$ ($d = 2$ ou 3) sont la projection d'une observation continue sur un sous espace de $\mathbb{R}^d$ muni d'une grille discrète de pixels dont la résolution dépend de l'instrument utilisé pour réaliser l'observation. L'ensemble des objets à détecter dans ces données $\mathbf{Y}$ peut être modélisé comme une configuration d'objets, *i.e.* comme une réalisation d'un processus ponctuel marqué. L'idée est de modéliser un objet qui peut parfois être complexe (en terme de forme, couleur, texture, etc) par un point : sa position dans l'image, et des marques simples qui vont caractériser cet objet. L'idée n'est pas de reconstruire complètement l'image à l'aide du processus, mais plutôt d'obtenir une modélisation assez simple d'une configuration d'objets dans une image. Ces marques, communes à tous les objets de la configuration, peuvent être vues comme un ensemble de variables aléatoires dont un objet est une réalisation. Pour définir le processus ponctuel marqué associé à la configuration d'objets que l'on cherche à estimer, il faut se poser les questions suivantes :

1. Quelle forme géométrique simple modélise au mieux l'ensemble des objets ?
2. Quelles autres marques sont nécessaires à la description des objets ?
3. Comment définir les interactions entre les objets d'une même configuration ?
4. Comment lier les caractéristiques des objets aux observations $\mathbf{Y}$?

B.3.1.1 Choix des objets

La première question est évidemment liée à l'application considérée. Le choix des marques des objets doit être guidé par la géométrie, et les caractéristiques générales des objets que nous cherchons à détecter dans l'image. Parmi les exemples cités précédemment, différents problèmes de détections d'objets ont été étudiés : dans les travaux de Ortner [2004], les objets recherchés sont des bâtiments qui sont alors modélisés par des parallélépipèdes rectangles ; un réseau routier a été modélisé par un ensemble de segments dans les travaux de Keresztes et al. [2009] ; et la détection de flamants roses à été réalisée à l'aide d'objets elliptiques dans les travaux de Descamps et al. [2008].

Le choix d'objets simples, pouvant être décrits entièrement par quelques paramètres, est important pour la phase de simulation du processus ponctuel marqué. Nous avons vu dans la partie B.2 que la simulation d'un tel processus passait par l'échantillonnage de la configuration d'objets en proposant des modifications sur le nombre, la forme, la position, etc, des objets.

Afin de limiter le coût calculatoire de l'estimation des marques des objets, il est important de réduire au minimum le nombre de descripteur de l'objet. Par exemple l'utilisation d'un cercle au lieu d'une ellipse lorsque cela est possible, réduit considérablement la dimension de l'espace des marques géométrique à explorer. En effet, un cercle est entièrement défini par un point : son centre, et son rayon (une position + une marque), tandis que l'ellipse nécessite la définition d'un point : son centre, la taille des deux demi-axes ainsi que son orientation (une position + 3 marques).

B.3.1.2 Définition de marques complexes

Outre les marques géométriques de l'objet, il peut parfois être nécessaire d'introduire d'autres marques particulières à l'application considérée. Nous verrons par exemple dans le chapitre 2, que toutes les sources à détecter présentent la même caractéristique spatiale : l'intensité de la source est maximale au centre et décroit en s'éloignant du bord. Dans ce cas, il est possible d'ajouter en plus des marques géométrique une marque caractérisant la décroissance spatiale d'intensité. Sur cet exemple, le profil d'intensité est défini de façon paramétrique par un indice n qui gère le taux de décroissance, et qui est différent d'une galaxie à l'autre. Cet indice n sera ajouté à l'ensemble des marques géométriques définissant les objets du processus ponctuel marqué.

B.3.1.3 Définition des interactions entre objets

Si la configuration d'objets à détecter dans une image possède des propriétés sur les interactions entre les objets, il est possible d'introduire cette connaissance *a priori* dans le modèle du processus ponctuel marqué afin de favoriser la simulation de configurations respectant les critères d'interactions. Il existe plusieurs types d'interactions entre les objets d'une configuration :

— la répulsion
— l'attraction,
— les recouvrements.

Par exemple, dans le cas des travaux de Keresztes et al. [2009] sur la détection de réseaux routiers dans des images de télédétection, les objets sont des segments modélisant des routes, et ces routes sont bien entendu reliées entre elles pour former le réseau. L'ajout d'un terme d'interaction de type attraction dans la densité du processus permet de favoriser les configurations où les segments sont reliés deux à deux et de pénaliser les configurations contenant au moins un segment isolé qui modéliserait alors une route commençant et ne menant nulle part. De même dans les travaux de Hadj et al. [2010], les objets à détecter sont des navires stationnés dans un port. La structure du port et les conventions d'amarrages des bateaux dans le port entraînent un alignement naturel des navires les uns à côtés des autres. Les configurations présentant des objets distribués de façon pêle-mêle dans l'image doivent donc être pénalisées par l'ajout d'un terme d'attraction entre les objets.

Les relations de type répulsion peuvent être utilisées pour éviter les détections multiples d'un même objet, les configurations contenant au moins deux objets qui se recouvrent partiellement seront alors pénalisées devant les configurations qui ne présentent aucune intersection entre objets. La relation de répulsion entre deux objets u_i et u_j la plus simple peut être proportionnelle à l'aire d'intersection de ces deux objets, voir par exemple dans les travaux de Descamps et al. [2008].

L'interdiction stricte de configurations contenant des objets qui se recouvrent trop (ou qui sont trop proches ou partagent une trop grande proportion de leur énergie selon les applications) est modélisée par un terme dit *hard-core* qui interdit les configurations contenant au moins une paire d'objets dont la proportion de recouvrement est supérieure à un seuil fixé tandis que les autres configurations ne sont ni pénalisées ni favorisées. Un exemple de pénalisation de type

hard-core est donné dans les travaux de Chatelain et al. [2009] où les configurations comportant au moins une paire d'objets partageant plus de 50% de leur aire sont interdites.

Il est également possible de combiner des interactions de types répulsion et attraction dans le même terme de pénalisation en définissant par exemple des zones géographiques autour de l'objet où sont favorisées les attractions et d'autres où sont favorisées les relations de répulsions. Un exemple peut être trouvé dans les travaux de Tournaire et al. [2007] pour la détection des marquages routiers rectangulaires blancs (lignes continues, lignes de dissuasion, bandes d'arrêt d'urgence, etc). Dans le cas d'une ligne pointillée séparant deux voies de circulation, les pointillés sont détectés par des objets rectangulaires qui sont régulièrement espacés : la distance entre deux rectangles d'une même ligne est constante ; il est donc naturel de favoriser les configurations de rectangles alignés et espacés de cette distance tout en pénalisant les configurations contenant par exemple des rectangles qui s'intersectent ou sont perpendiculaires deux à deux.

B.3.1.4 Lien entre la configuration d'objet et les données

Le lien entre la configuration d'objets modélisée par le processus ponctuel marqué et les données se traduit par la présence d'un terme d'attache aux données dans la densité du processus. Ce terme d'attache aux données peut être une densité obtenue par la vraisemblance des données comme nous allons le voir dans le paragraphe B.3.2 ou un terme plus complexe obtenu comme la somme des attaches locales des objets aux données.

B.3.2 Estimer la configuration d'objets à partir de sa densité

Pour estimer une configuration d'objets dans des données $\mathbf{Y}$, nous utilisons un processus ponctuel marqué caractérisé par sa densité $p(\boldsymbol{u}|\mathbf{Y})$ exprimée conditionnellement aux données. L'ensemble des objets à détecter dans ces données $\mathbf{Y}$ peut être modélisé comme la configuration $\hat{\boldsymbol{u}}$ maximisant la densité $p(\boldsymbol{u}|\mathbf{Y})$ d'un processus ponctuel marqué :

$$\widehat{\boldsymbol{u}} = \underset{\boldsymbol{u}\in\Omega}{\operatorname{argmax}} \Big\{ p(\boldsymbol{u}|\mathbf{Y}) \Big\} \tag{B.17}$$

Lorsque les données peuvent être décrites par un modèle paramétrique conditionnellement à une configuration d'objets donnée, il est possible de définir la fonction de vraisemblance $\mathcal{L}(\boldsymbol{u}, \mathbf{Y})$ des observations et la configuration $\hat{\boldsymbol{u}}$ peut être estimée au sens du maximum de vraisemblance (MV). Dans ce cas, $p(\boldsymbol{u}|\mathbf{Y}) = \mathcal{L}(\boldsymbol{u}, \mathbf{Y})$. Dans un cadre bayésien, la densité $p(\boldsymbol{u}|\mathbf{Y})$ peut s'écrire comme le produit d'une densité *a priori* $\pi(\boldsymbol{u})$ sur la configuration d'objets $\boldsymbol{u}$ et de la fonction de vraisemblance $\mathcal{L}(\boldsymbol{u}, \mathbf{Y})$ qui relie la configuration d'objets aux observations $\mathbf{Y}$. L'estimateur maximisant la densité $p(\boldsymbol{u}|\mathbf{Y})$ est alors appelé estimateur au sens du *maximum a posteriori* (MAP).

B.3.2.1 Estimateur au sens du maximum de vraisemblance *vs* Estimateur au sens du maximum *a posteriori*

Pour les problématiques de détection d'objets dans une image, l'estimateur de la configuration d'objets au sens du *maximum a posteriori* est souvent préféré à l'estimateur au sens du maximum de vraisemblance. L'un des arguments avancés par Baddeley and Van Lieshout [1993] est que l'estimateur MV converge vers une solution contenant souvent des détections multiples, *i.e.* une même source est détectée par plusieurs objets qui se recouvrent partiellement, afin d'expliquer au mieux les données. Ce phénomène est particulièrement gênant lorsque l'un des objectifs de la détection est d'estimer le nombre d'objets présents dans l'image.

L'approche bayésienne associée a l'estimateur MAP de la configuration d'objets permet d'inclure des contraintes sur le recouvrement entre les objets, sur les relations de voisinage, etc. Il

est notamment possible d'interdire le recouvrement total ou quasi total de deux objets. L'estimateur MAP de la configuration est alors obtenue par optimisation stochastique sur l'espace de l'ensemble des configurations Ω :

$$\widehat{\boldsymbol{u}} = \underset{\boldsymbol{u}\in\Omega}{\operatorname{argmax}}\Big\{\mathcal{L}(\boldsymbol{u},\mathbf{Y})\pi(\boldsymbol{u})\Big\} \Leftrightarrow \underset{\boldsymbol{u}\in\Omega}{\operatorname{argmax}}\Big\{\log(\mathcal{L}(\boldsymbol{u},\mathbf{Y})) + \log(\pi(\boldsymbol{u}))\Big\} \tag{B.18}$$

où $\mathbf{Y}$ représente les observations, $\mathcal{L}(\boldsymbol{u},\mathbf{Y})$ est la fonction de vraisemblance des observations et $\pi(\boldsymbol{u})$ l'a priori placé sur la configuration d'objet $\boldsymbol{u}$. Le terme $\log(\mathcal{L}(\boldsymbol{u},\mathbf{Y}))$ peut-être identifié comme le terme d'attache aux données et $\log(\pi(\boldsymbol{u}))$ est une contrainte de pénalisation sur les configurations d'objets.

L'estimation de la configuration d'objets peut être réalisée conjointement à l'estimation d'autres paramètres du modèle (par exemple des paramètres de moyenne et de variance liés à la modélisation du bruit). Il suffit pour cela de mélanger les différents algorithmes d'échantillonnage présentés dans le paragraphe B.2, ce sera notamment le cas de la méthode de détection et d'estimation présentée dans le chapitre 2 de ce manuscrit.

B.3.2.2 Estimation de la configuration avec une formulation énergétique

De nombreux auteurs dans la littérature préfèrent une formulation énergétique de la densité du processus ponctuels marqués, voir par exemple [Descombes, 2011, chapitre 5], Descamps et al. [2008], ou Stoica et al. [2004]. Dans ce cas, la densité $p(\boldsymbol{u}|\mathbf{Y})$ s'écrit :

$$p(\boldsymbol{u}|\mathbf{Y}) = \frac{\exp\left(-U(\boldsymbol{u}|\boldsymbol{\theta})\right)}{c(\boldsymbol{\theta})},$$

où $U(\boldsymbol{u}|\boldsymbol{\theta})$ est l'énergie du système conditionnellement aux paramètres qui définissent les marques du processus ponctuel marqué, et $c(\boldsymbol{\theta})$ est une constante de normalisation. Cette modélisation sous forme d'énergie peut être interprétée de la façon suivante : lors de l'échantillonnage d'un processus, l'étape d'acceptation-rejet consiste en fait à évaluer l'énergie nécessaire à l'ajout, la suppression ou la modification d'un objet. Cette énergie s'écrit : $\log(\ p(\boldsymbol{v}|\mathbf{Y})\ ) - \log(\ p(\boldsymbol{u}|\mathbf{Y})\ )$ où $\boldsymbol{u}$ est la configuration courante et $\boldsymbol{v}$ est la configuration proposée. Le terme d'énergie $U(\boldsymbol{u}|\boldsymbol{\theta})$ peut être décomposé en la somme de deux termes :

$$U(\boldsymbol{u}|\boldsymbol{\theta}) = U_{\mathbf{Y}}(\boldsymbol{u}|\boldsymbol{\theta}) + U_i(\boldsymbol{u}|\boldsymbol{\theta}),$$

où $U_{\mathbf{Y}}(\boldsymbol{u}|\boldsymbol{\theta})$ est un terme d'attache aux données et $U_i(\boldsymbol{u}|\boldsymbol{\theta})$ est un terme d'énergie d'interaction entre les objets de la configuration. La configuration d'objets qui maximise $p(\boldsymbol{u}|\mathbf{Y})$ est équivalente à :

$$\widehat{\boldsymbol{u}} = \underset{\boldsymbol{u}\in\Omega}{\operatorname{argmin}}\Big\{U_{\mathbf{Y}}(\boldsymbol{u}|\boldsymbol{\theta}) + U_i(\boldsymbol{u}|\boldsymbol{\theta})\Big\}$$

Sous cette formulation, $\widehat{\boldsymbol{u}}$ peut être interprétée comme un estimateur au sens du maximum *a posteriori* en identifiant :

- $U_{\mathbf{Y}}(\boldsymbol{u}|\boldsymbol{\theta})$ à l'opposée de la log-vraisemblance : $-\log(\mathcal{L}(\boldsymbol{u},\mathbf{Y}))$ introduite dans le paragraphe B.3.2.1,
- $U_i(\boldsymbol{u}|\boldsymbol{\theta})$ à $-\log(\pi(\boldsymbol{u}))$ qui est le logarithme de la loi *a priori* sur les objets.

L'énergie d'interaction peut par exemple résumer la contrainte de connectivité et d'alignement des segments dans la détection des réseaux routiers.

Annexe C

Modélisation du profil d'intensité des galaxies par un profil Sersic

Dans cette annexe, nous présentons le calcul du profil d'intensité des objets du processus ponctuel marqué utilisé pour modéliser la configuration de galaxies. Le choix du profil d'intensité s'est porté sur les fonctions Sersic (Sersic [1963]) largement utilisées dans la littérature astrophysique pour représenter la décroissance d'intensité centre-bord des galaxies (voir par exemple Trujillo et al. [2001], Simard et al. [2002], Perret et al. [2009]). Nous utilisons ici un profil Sersic définit en deux dimensions sur un support elliptique modélisant la forme de la galaxie. L'intérêt de la méthode de détection proposée est d'utiliser une approche objet, ce qui permet d'effectuer des traitements localement plutôt que de considérer le cube de données dans son ensemble (soit plus de 324 millions de pixels).

C.1 Cahier des charges

La modélisation d'une galaxie par un objet u_i passe tout d'abord par la définition d'un profil d'intensité spatial Sersic qui sera défini sur un support elliptique. Afin de définir parfaitement le support elliptique quelque soit le profil Sersic utilisé, il faut définir quelques règles :

1. le centre est défini de manière continue dans l'image, cependant le profil d'intensité sera ensuite échantillonné selon la grille de pixels de l'image,
2. il n'y a pas de distinction petit axe et grand axe,
3. l'indice Sersic n est donné,
4. la largeur à mi-hauteur du profil Sersic est donnée au niveau des deux axes de l'ellipse,
5. le profil au niveau des deux axes de l'ellipse doit contenir 95% de l'énergie du profil.

Il faut noter que le profil Sersic définit par l'équation (2.1) dont on rappelle la forme ici :

$$I(r) = I_0 \exp\left(-r^{\frac{1}{n}}\right),$$

est défini en fonction d'une variable $r = r(x, y)$ qui décrit chaque position pixellique (x, y) dans le repère elliptique lié à l'objet que l'on souhaite décrire. Nous allons définir les transformations nécessaires à la création de ce repère elliptique à partir du repère cartésien associé aux dimensions spatiales du cube de données.

C.2 Caractérisation du repère elliptique

Considérons ici le cas général d'une ellipse centrée aux coordonnées (x_0, y_0) dont les demi-axes mesurent β_1 et β_2 dans un repère cartésien $\mathcal{R} = (O, \overrightarrow{u}, \overrightarrow{v})$, où $\overrightarrow{u}$ et $\overrightarrow{v}$ sont les vecteurs

unitaires associés aux deux dimensions spatiales du cube de données. Le premier axe de l'ellipse forme un angle α avec l'axe des abscisses (O, $\overrightarrow{u}$). Notons $\mathcal{R}' = (O', \overrightarrow{u'}, \overrightarrow{v'})$ le repère cartésien où $\overrightarrow{u'}$ et $\overrightarrow{v'}$ sont les vecteurs unitaires associés aux axes de l'ellipse et O' est le centre de l'ellipse. Le repère $\mathcal{R}'$ est obtenue après translation et rotation du repère $\mathcal{R}$. On a donc :

$$\begin{cases} \overrightarrow{u'} &= cos(\alpha)\overrightarrow{u} + sin(\alpha)\overrightarrow{v} \\ \overrightarrow{v'} &= -sin(\alpha)\overrightarrow{u} + cos(\alpha)\overrightarrow{v} \end{cases},$$

et donc pour le point de coordonnées (x, y) dans $\mathcal{R} = (O, \overrightarrow{u}, \overrightarrow{v})$ et de coordonnées (x', y') dans $\mathcal{R}' = (O', \overrightarrow{u'}, \overrightarrow{v'})$ on a la relation suivante :

$$\begin{cases} x' &= cos(\alpha)(x - x_0) &+ sin(\alpha)(y - y_0) \\ y' &= -sin(\alpha)(x - x_0) &+ cos(\alpha)(y - y_0) \end{cases}$$

Le repère elliptique $\tilde{\mathcal{R}}$ associé à l'ellipse considérée est ensuite obtenu par homothétie sur les deux axes du repère $\mathcal{R}'$:

$$\begin{cases} \overrightarrow{\tilde{u}} &= \beta_1 \overrightarrow{u'} \\ \overrightarrow{\tilde{v}} &= \beta_2 \overrightarrow{v'} \end{cases}$$

où β_1 (respectivement β_2) représente la longueur du premier (respectivement du second) demi-axe de l'ellipse. Ainsi pour passer du repère $\mathcal{R}$ au repère $\tilde{\mathcal{R}}$ il faut effectuer les transformations suivantes :

$$\begin{cases} \tilde{x} &= \frac{cos(\alpha)}{\beta_1}(x - x_0) &+ \frac{sin(\alpha)}{\beta_1}(y - y_0) \\ \tilde{y} &= -\frac{sin(\alpha)}{\beta_2}(x - x_0) &+ \frac{cos(\alpha)}{\beta_2}(y - y_0) \end{cases}$$

La variable r de l'équation (2.1) qui représente la distance au centre de la galaxie dans le repère elliptique $\tilde{\mathcal{R}}$ s'écrit en fonction des coordonnées (x, y) dans le repère $\mathcal{R}$ de l'image de la façon suivante :

$$\begin{aligned} r^2 &= \tilde{x}^2 + \tilde{y}^2 \\ &= \frac{cos^2(\alpha)}{\beta_1^2}(x - x_0)^2 + \frac{sin^2(\alpha)}{\beta_1^2}(y - y_0)^2 + \frac{2cos(\alpha)sin(\alpha)}{\beta_1^2}(x - x_0)(y - y_0) \\ &\quad + \frac{sin^2(\alpha)}{\beta_2^2}(x - x_0)^2 + \frac{cos^2(\alpha)}{\beta_2^2}(y - y_0)^2 - \frac{2cos(\alpha)sin(\alpha)}{\beta_2^2}(x - x_0)(y - y_0) \end{aligned} \quad \text{(C.1)}$$

La figure C.1 illustre la construction du repère elliptique $\tilde{\mathcal{R}}$ à partir d'une ellipse caractérisée dans le repère cartésien $\mathcal{R}$.

C.3 Modélisation de l'objet dans ce repère elliptique

On cherche maintenant à modéliser une galaxie par un profil Sersic d'indice n en deux dimensions défini sur un support elliptique de centre (x_0, y_0) de demi-axes axe_1 et axe_2 et d'orientation α par rapport à l'axe des abscisses du repère cartésien $\mathcal{R}$ associé au cube de données MUSE. Cet objet doit respecter les contraintes énoncées dans le paragraphe C.1. D'après les conditions 3 et 4, l'indice Sersic n est fixé pour le profil en deux dimensions et les largeurs à mi-hauteur, $FWHM_1$ et $FWHM_2$, sont connues au niveau des axes de l'ellipse. Il faut donc déterminer les paramètres d'échelle β_1 et β_2 du repère elliptique $\tilde{\mathcal{R}}$ afin de l'ajuster à l'objet à modéliser. La figure C.2 illustre le problème à résoudre.

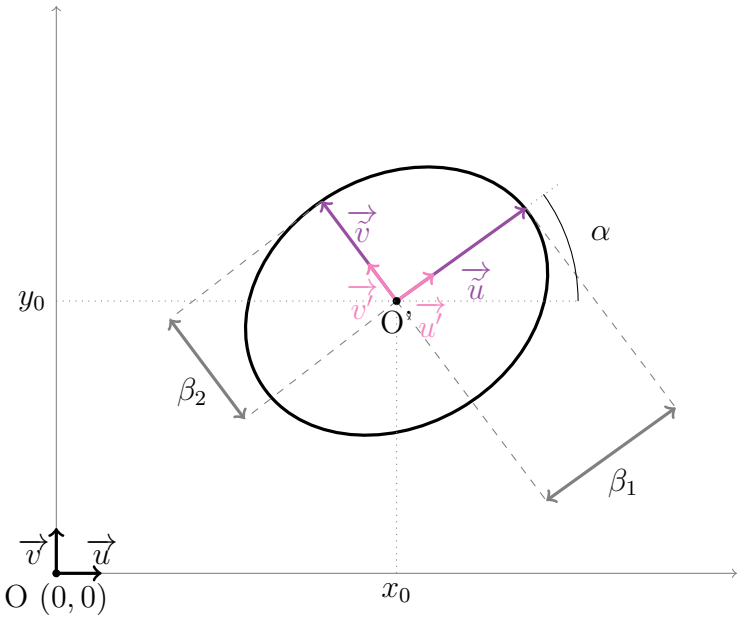


FIGURE C.1 – Définition d'un repère elliptique $\tilde{\mathcal{R}}$ à partir d'une ellipse caractérisée par son centre (x_0, y_0), ses demi-axes, β_1 et β_2 et son orientation α dans le repère cartésien $\mathcal{R}$

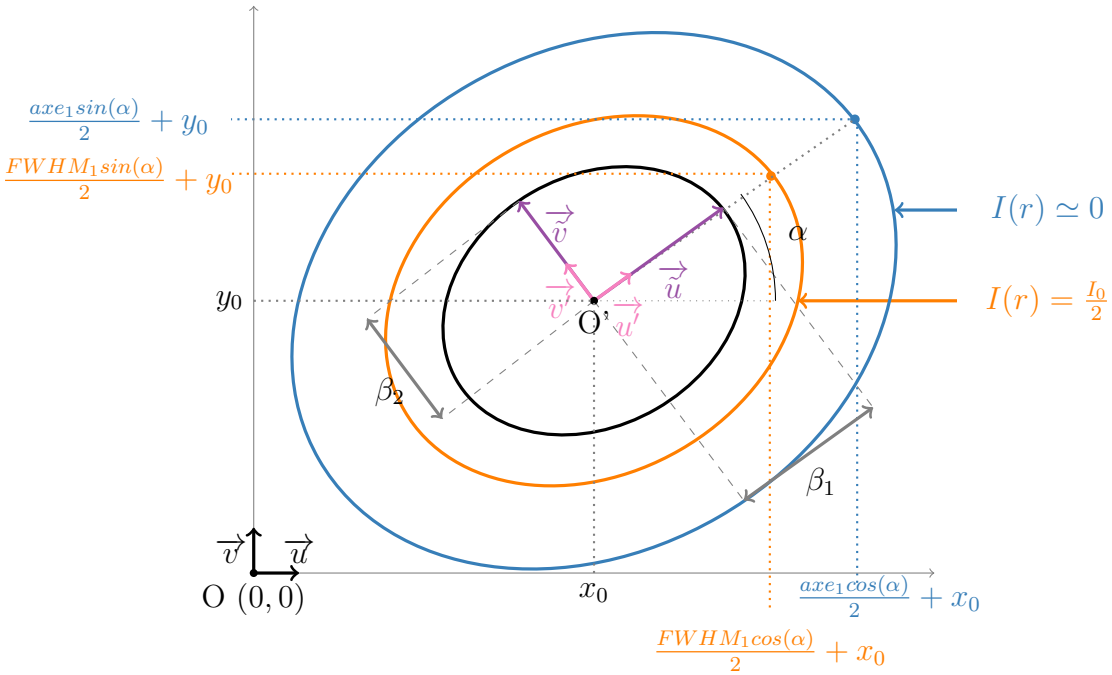


FIGURE C.2 – Représentation du support elliptique associé à un profil Sersic d'indice n et dont les largeurs à mi-hauteur $FWHM_1$ et $FWHM_2$ sont fixées. Le contour noir est l'ellipse qui définit le repère elliptique $\tilde{\mathcal{R}}$ associé à l'objet que l'on souhaite modéliser, il permet de respecter les conditions énoncées dans le paragraphe C.1. L'ellipse orange correspond à l'ensemble des points pour lesquels le profil Sersic en deux dimensions atteint la moitié de son intensité maximale, et l'ellipse bleue modélise la troncature du profil Sersic afin de contenir 95% de l'énergie du profil.

Si l'on se place sur l'axe_1 de l'objet elliptique que l'on cherche à modéliser. On connait l'intensité du profil Sersic au point de coordonnées $r(x,y) = FWHM_1/2$, elle doit valoir la moitié de l'intensité au centre de l'ellipse. De même si l'on se place au niveau de l'$axe2$. Ces deux conditions peuvent se traduire par :

$$\begin{cases} \exp\left(-\left(\frac{FWHM_1}{2\beta_1}\right)^{\frac{1}{n}}\right) &= \frac{1}{2} \\ \exp\left(-\left(\frac{FWHM_2}{2\beta_2}\right)^{\frac{1}{n}}\right) &= \frac{1}{2} \end{cases}$$

Finalement les paramètres β_1 et β_2 s'écrivent :

$$\begin{cases} \beta_1 &= \frac{FWHM_1}{2ln(2)^n} \\ \beta_2 &= \frac{FWHM_2}{2ln(2)^n} \end{cases}$$

Afin de décrire le profil d'intensité Sersic en deux dimensions, nous allons donc définir un repère elliptique associé à l'ellipse dont la longueur des demi-axes est respectivement β_1 et β_2, le centre et l'orientation sont les mêmes que ceux de l'objet elliptique u_i modélisant la galaxie considérée. Nommons axe_1 et axe_2 les demi-axes de l'objet u_i, la longueur de ces axes est définie grâce à la condition 5 qui spécifie que le support elliptique associé au profil spatial d'intensité doit contenir 95% de l'énergie du profil Sersic au niveau des axes. Si nous nous plaçons dans la direction de l'axe_1 (respectivement de l'axe_2), on veut respecter la condition suivante :

$$\int_{-R_{max}}^{R_{max}} I(|r|)dr = 0,95 \int_{-\infty}^{\infty} I(|r|)dr \Leftrightarrow \int_0^{R_{max}} \exp\left(-r^{\frac{1}{n}}\right)dr = 0,95 \int_0^{\infty} \exp\left(-r^{\frac{1}{n}}\right)dr, \quad \text{(C.2)}$$

où $R_{max} = axe_1/\beta_1$ (respectivement $R_{max} = axe_2/\beta_2$) dans la direction de l'axe_1 (respectivement de l'axe_2). Posons

$$\begin{cases} t &= r^{\frac{1}{n}} \\ dt &= \frac{1}{n} r^{\frac{1}{n}-1} dr \end{cases} \Leftrightarrow \begin{cases} r &= t^n \\ dr &= nt^{n-1}dt \end{cases}$$

L'équation (C.2) peut se réécrire :

$$n \int_0^{R_{max}^{\frac{1}{n}}} t^{n-1} \exp(-u)du = 0,95n \int_0^{\infty} t^{n-1} \exp(-u)du \quad \text{(C.3)}$$

On reconnait dans le membre de gauche la fonction gamma incomplète :

$$\gamma(\cdot, x) : n \mapsto \int_0^x nt^{n-1} \exp(-u)du,$$

et dans le membre de droite la fonction gamma

$$\Gamma : n \mapsto \int_0^{\infty} nt^{n-1} \exp(-u)du.$$

Ainsi l'équation (C.3) se réécrit après simplifcation :

$$\gamma(R_{max}^{\frac{1}{n}}, x) = 0,95\Gamma(n) \quad \text{(C.4)}$$

La fonction gamma incomplète γ et son inverse γ_{inv} sont implémentées dans la plupart des langages scientifiques, il suffit alors pour déterminer axe_1 (respectivement axe_2) de calculer $R_{max} = axe_i/\beta_i = \gamma_{inv}(n, 0,95\Gamma(n))^n$.

D'un point de vue de l'implémentation, nous avons ajouter une contrainte supplémentaire sur la taille maximale des axes de l'ellipse support. Si la longueur du demi-axe obtenu est supérieure à quatre fois la demi largeur à mi-hauteur associée, alors l'axe est tronqué à cette valeur. En effet, pour des indices Sersic $n \geqslant 2$ et une largeur à mi-hauteur assez grande, le profil décroit lentement, et pour contenir 95% de l'énergie du profil, le support elliptique pourrait s'étendre sur plusieurs dizaines de pixels de diamètre. Or l'algorithme est contraint par le taux de recouvrement entre les objets de la configuration. Un fort taux de recouvrement entre les objets entrainerait une augmentation de la durée d'exécution de l'algorithme (notamment dans le calcul et la mise à jour des termes faisant intervenir l'inverse de la matrice de Gram $\mathbf{X}^T\mathbf{X}$ qui modélise les interactions entre les objets). De plus d'un point de vue physique, dans les champs profonds les galaxies observées n'ont pas une extension spatiale infinie, même après convolution par la FSF, le support sur lequel l'intensité de la galaxie est significative est limité dans l'espace.

C.4 Modélisation du profil Sersic en deux dimensions

Les figures présentées dans ce paragraphe (figure C.3) correspondent aux différents profils Sersic et le support qui leur est associé pour les contraintes suivantes :

- le centre a pour coordonnées dans le repère $\mathcal{R}$: $(x_0, y_0) = (15, 15)$
- les largeurs à mi-hauteur sont : $FWHM_1 = 3$ et $FWHM_2 = 4$ pixels
- l'orientation de l'ellipse dans le repère $\mathcal{R}$ est $\alpha = \pi/3$

Les profils Sersic englobent les profils gaussiens et exponentiels, nous retrouvons un profil gaussien en deux dimensions pour l'indice $n = 0.5$, les constantes β_i sont équivalentes à $\sqrt{2}\sigma_i$ où σ_i^2 est la variance de la gaussienne définie sur l'axe_i. L'indice $n = 1$ correspond à une décroissance d'intensité exponentielle où les β_i représentent simplement des facteurs d'échelle.

Pour une même largeur à mi-hauteur, le profil Sersic devient de plus en plus piqué lorsque n augmente, et la décroissance d'intensité est d'autant plus lente. Alors que les profils correspondant aux indices $n = 0.5$ et $n = 1$ montrent une décroissance vers zéro aux bords du support elliptique, le profil obtenu avec l'indice $n = 2$ illustre la troncature réalisée afin de limiter l'extension spatiale du support. Pour $FWHM_1 = 3$ on obtient $\beta_1 = \frac{FWHM_1}{2ln(2)^n} = \frac{3}{2ln(2)^2} = 3.12$ et donc $axe_1 = \gamma_{inv}(n, 0.95\Gamma(n))^n\beta_1 = 70.26$ pixels. Afin de respecter la condition 5 de préservation de l'énergie, cette galaxie serait donc modéliser par un profil Sersic associé à un support elliptique dont l'extension selon le premier axe serait de 140 pixels et selon le deuxième axe de 187 pixels (à comparer aux dimensions spatiales du cube de données de 300×300 pixels). Avec cet exemple, nous illustrons la nécessité de tronquer le support elliptique de l'objet dans certaines conditions afin de maintenir une modélisation réaliste des données. La troncature sera cependant adoucie lors de la convolution avec la FSF moyenne (voir équation (2.6), où s représente le profil Sersic en deux dimensions et F la FSF moyenne).

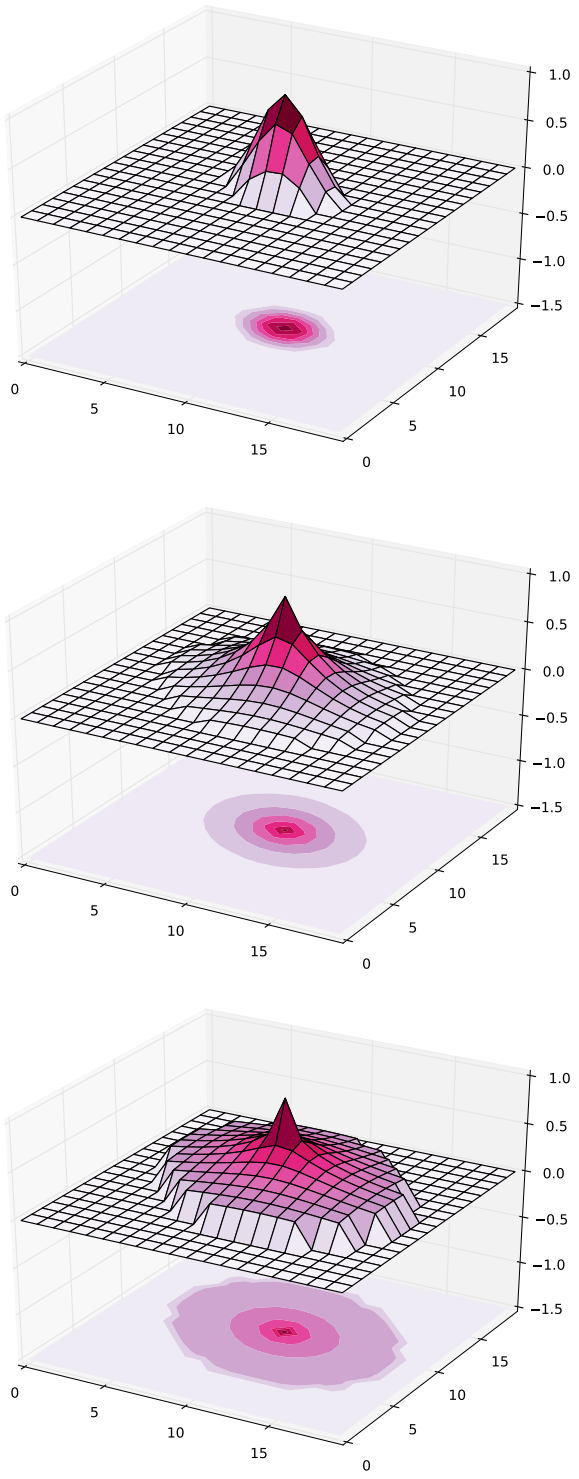


Figure C.3 – Profils Sersic et supports elliptiques associés définis selon les mêmes contraintes de centre, de largeurs à mi-hauteur et d'orientation pour différents indices n. Haut : $n = 0.5$, centre : $n = 1$ et bas = $n = 2$.

Annexe D

Mise à jour récursive de la matrice de Gram

Dans cette annexe, nous considérons un mouvement sur la configuration d'objet entrainant un changement de dimension un de la matrice de Gram $\mathbf{X}^T\mathbf{X}$. Nous nous intéressons donc aux mouvements de type naissance-mort pour l'échantillonneur de Metropolis-Hastings-Green. Pour l'ajout d'un objet, la configuration de n objets passe à une configuration de $n-1$ objets. De même pour la suppression d'un objet, la configuration verra son nombre d'objets décroître de n à $n-1$. Les notations utilisées ici sont : $\mathbf{X}$ la matrice de taille $M \times n$ décrivant la configuration courante et $\widetilde{\mathbf{X}}$ la matrice représentant la nouvelle configuration. Par simplicité nous ne considérerons ici que le cas où la matrice de configuration $\mathbf{X}$ ne varie pas avec la longueur d'onde.

Dans l'algorithme d'échantillonnage RJMCMC, nous sommes amenés à évaluer la matrice de Gram $\mathbf{X}^T\mathbf{X}$, ainsi que son inverse et sa multiplication avec les données. Lors de chaque mouvement sur la configuration d'objet, il faut recalculer l'ensemble de ces quantités. Or, le calcul de l'inverse de la matrice de Gram $\mathbf{X}^T\mathbf{X}$, ou de sa décomposition de Cholesky $\mathbf{C}\mathbf{C}^T = \mathbf{X}^T\mathbf{X}$ où $\mathbf{C}$ est triangulaire inférieure, présente une complexité en $\mathcal{O}(n^3)$. Ce calcul, dont la complexité dépend du nombre d'objets détectés est trop coûteuse à évaluer dans le cas où la configuration contient un grand nombre d'objets. C'est pourquoi nous avons développé une stratégie de mise à jour récursive sur la dimension de cette matrice de Cholesky $\mathbf{C}$ présentant une complexité en $\mathcal{O}(n^2)$.

Nous détaillerons dans cette annexe la mise à jour récursive des matrices suivantes :

— $\widetilde{C}$, qui correspond à la décomposition de Cholesky de la matrice de Gram de la nouvelle configuration $\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}$, avec $\widetilde{C}$ qui est triangulaire inférieure,

— $\widetilde{C}^{-1}\widetilde{\mathbf{X}}^T[\mathbf{1}, \boldsymbol{y}_1, ..., \boldsymbol{y}_\Lambda]$, qui apparait dans l'expression de la densité à posteriori de la nouvelle configuration. Il faut noter que les produits avec toutes les longueurs d'onde sont mis à jour dans ce terme.

D.1 Mouvement de naissance

Le mouvement de naissance consiste à ajouter un objet à la configuration courante, ce qui se traduit au niveau de la modélisation matricielle par l'ajout d'une colonne à la matrice de configuration :

$$\mathbf{X} = \left(\begin{array}{c|c|c} \vdots & & \vdots \\ \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_n \\ \vdots & & \vdots \end{array} \right)$$

qui devient :

$$\widetilde{\mathbf{X}} = \left(\begin{array}{c|c|c|c} \vdots & & \vdots & \vdots \\ \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_n & \boldsymbol{x}_{n+1} \\ \vdots & & \vdots & \vdots \end{array} \right),$$

et

$$\widetilde{C}\widetilde{C}^T = \widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}} = \left(\begin{array}{c|c} CC^T & \mathbf{X}^T\boldsymbol{x}_{n+1} \\ \hline \boldsymbol{x}_{n+1}^T\mathbf{X} & \boldsymbol{x}_{n+1}^T\boldsymbol{x}_{n+1} \end{array} \right)$$

Le calcul du vecteur $\boldsymbol{x}_{n+1}^T\mathbf{X}$ et du produit scalaire $\boldsymbol{x}_{n+1}^T\boldsymbol{x}_{n+1}$ est peu coûteux en terme de calcul puisque $\boldsymbol{x}_{n+1}$ et $\mathbf{X}$ présentent des structures creuses (la taille du support spatial des objets est très faible devant les dimensions spatiales des données). Le même commentaire peut s'appliquer au produit scalaire de la matrice $\mathbf{X}$ avec les données $[\mathbf{1}, \boldsymbol{y}_1, ..., \boldsymbol{y}_\Lambda]$.

Algorithme de mise à jour récursive :

1. $v = \boldsymbol{x}_{n+1}^T\boldsymbol{x}_{n+1}$
2. $\boldsymbol{v_p} = C^{-1}\mathbf{X}^T\boldsymbol{x}_{n+1}$
3. $g = \sqrt{v - \boldsymbol{v_p}^T\boldsymbol{v_p}}$
4. $\widetilde{C} = \left(\begin{array}{c|c} C & \begin{matrix} 0 \\ \vdots \\ 0 \end{matrix} \\ \hline \boldsymbol{v_p}^T & g \end{array} \right)$
5. $$\widetilde{C}^{-1}\widetilde{\mathbf{X}}^T[\mathbf{1}, \boldsymbol{y}_1, ..., \boldsymbol{y}_\Lambda] = \left(\begin{array}{c} C^{-1}\mathbf{X}^T[\mathbf{1}, \boldsymbol{y}_1, ..., \boldsymbol{y}_\Lambda] \\ \hline \frac{1}{g}\left(\boldsymbol{x}_{n+1} - \boldsymbol{v_p}^T C^{-1}\mathbf{X}^T\right)[\mathbf{1}, \boldsymbol{y}_1, ..., \boldsymbol{y}_\Lambda] \end{array} \right)$$

   On introduit $\boldsymbol{a} = \frac{1}{g}\left(x_{n+1} - v_p^T C^{-1}\mathbf{X}^T\right)[\mathbf{1}, \boldsymbol{y}_1, ..., \boldsymbol{y}_\Lambda]$, ce qui permet d'écrire :

$$\widetilde{C}^{-1}\widetilde{\mathbf{X}}^T[\mathbf{1}, \boldsymbol{y}_1, ..., \boldsymbol{y}_\Lambda] = \left(\begin{array}{c} C^{-1}\mathbf{X}^T[\mathbf{1}, \boldsymbol{y}_1, ..., \boldsymbol{y}_\Lambda] \\ \hline \boldsymbol{a} \end{array} \right) = \left(\begin{array}{c} C^{-1}\mathbf{X}^T[\mathbf{1}, \boldsymbol{y}_1, ..., \boldsymbol{y}_\Lambda] \\ \hline \mathbf{0} \end{array} \right) + \left(\begin{array}{c} \mathbf{0} \\ \hline \boldsymbol{a} \end{array} \right)$$

Finalement en développant le calcul de $[\mathbf{1}, \boldsymbol{y}_1, ..., \boldsymbol{y}_\Lambda]^T\widetilde{\mathbf{X}}(\widetilde{C}\widetilde{C}^T)^{-1}\widetilde{\mathbf{X}}^T[\mathbf{1}, \boldsymbol{y}_1, ..., \boldsymbol{y}_\Lambda]$, on obtient :

$$[\mathbf{1}, \boldsymbol{y}_1, ..., \boldsymbol{y}_\Lambda]^T\widetilde{\mathbf{X}}(\widetilde{C}\widetilde{C}^T)^{-1}\widetilde{\mathbf{X}}^T[\mathbf{1}, \boldsymbol{y}_1, ..., \boldsymbol{y}_\Lambda] = [\mathbf{1}, \boldsymbol{y}_1, ..., \boldsymbol{y}_\Lambda]^T\mathbf{X}(CC^T)^{-1}\mathbf{X}^T[\mathbf{1}, \boldsymbol{y}_1, ..., \boldsymbol{y}_\Lambda] + \boldsymbol{a}^T\boldsymbol{a}$$

Le ratio des déterminants $\left(\frac{\det \widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}}{\det \mathbf{X}^T\mathbf{X}}\right)$ présent dans le ratio de Métropolis-Hastings-Green pour une naissance se simplifie par :

$$\left(\frac{\det \widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}}{\det \mathbf{X}^T\mathbf{X}}\right) = \left(\frac{\det \widetilde{\mathbf{C}}\widetilde{\mathbf{C}}^T}{\det \mathbf{C}\mathbf{C}^T}\right) = \left(\frac{\left(\det \widetilde{\mathbf{C}}\right)^2}{(\det \mathbf{C})^2}\right) = \left(\frac{\det \mathbf{C} \times g}{\det \mathbf{C}}\right)^2 = g^2.$$

On peut remarquer que lorsque le nouvel objet est orthogonal aux autres objets de la configuration (*i.e.* en l'absence de recouvrement de la réponse de l'objet avec celles de ceux déjà présents),

le ratio des déterminants vaut 1. En effet dans ce cas $g = \sqrt{v - \boldsymbol{v_p}^T \boldsymbol{v_p}} = \sqrt{v} = \sqrt{\boldsymbol{x}_{n+1}^T \boldsymbol{x}_{n+1}} = 1$ car, par convention, les réponses des objets sont normalisées. Remarquons également que le ratio des déterminants est borné. Par construction g est inférieur ou égal à 1. De plus, g ne peut pas tendre vers zéro du fait de la présence de la régularisation hard core définie dans l'équation (2.21) qui pénalise le recouvrement entre les objets au delà de la limite du critère de Rayleigh. Le produit scalaire de la réponse $\boldsymbol{x}_{n+1}$ du nouvel objet avec les réponses des objets déjà présents $\boldsymbol{x}_i$, $i = 1, \cdots, n$ ne peut être supérieur à la valeur t définie dans l'équation (2.21), $0 < t < 1$. Ceci assure que la loi *a posteriori*, et donc le ratio de Metropolis-Hastings-Green, sont bien définis.
Le calcul du ratio de Metropolis-Hastings-Green donné dans l'équation (2.29) qui dépend uniquement de la dernière ligne de la matrice $\widetilde{C}^{-1}\widetilde{\mathbf{X}}^T [\mathbf{1}, \boldsymbol{y}_1, ..., \boldsymbol{y}_\Lambda]$, devient :

$$r(\boldsymbol{u}, \boldsymbol{v}) = q \times g^{-\Lambda} \times \prod_{\lambda=1}^{\Lambda} \exp\left\{ \frac{a_{\lambda+1}^2 - 2m_\lambda a_1 a_{\lambda+1} + m_\lambda^2 a_1^2}{2\sigma_\lambda^2} \right\} \times \sqrt{2\pi\sigma_\lambda^2}$$

où a_i est la $i^{\text{ème}}$ composante du vecteur ligne $\boldsymbol{a}$ de taille $(\Lambda + 1)$.

D.2 Mouvement de mort d'un objet

Le mouvement de mort consiste à supprimer une colonne de la matrice $\mathbf{X}$:

$$\mathbf{X} = \left(\begin{array}{c|c|c|c|c} \vdots & & \vdots & & \vdots \\ \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_j & \cdots & \boldsymbol{x}_n \\ \vdots & & \vdots & & \vdots \end{array} \right)$$

becomes :

$$\widetilde{\mathbf{X}} = \left(\begin{array}{c|c|c|c|c} \vdots & & \vdots & \vdots & \vdots \\ \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_{j-1} & \boldsymbol{x}_{j+1} & \boldsymbol{x}_n \\ \vdots & & \vdots & \vdots & \vdots \end{array} \right)$$

La suppression de la $j^{\text{ème}}$ ligne et de la $j^{\text{ème}}$ colonne de la matrice de Gram $\mathbf{X}^T\mathbf{X}$ nécessite la mise à jour de la décomposition de Cholesky C. Pour cela, la première étape consiste à appliquer des transformations aux différentes matrices à mettre à jour afin de placer les informations relative à l'objet à supprimer sur la dernière ligne et/ou dernière colonne. Il faut ensuite trianguliser la matrice de Cholesky ainsi modifiée et supprimer la $j^{\text{ème}}$ ligne et la dernière colonne pour obtenir la nouvelle matrice $\widetilde{C}$ de taille $n - 1 \times n - 1$. Soit $j \in \{1, \ldots, n\}$ l'indice dans la matrice $\mathbf{X}$ de l'objet à supprimer. Par la suite, nous utiliserons les notations matricielles suivantes :

- $A_{i,j}$ est l'élément de la $i^{\text{ème}}$ ligne et $j^{\text{ème}}$ colonne de la matrice A,
- A_l désigne la $l^{\text{ème}}$ ligne de la matrice A,
- $A_{\backslash l}$ la matrice A privée de sa $l^{\text{ème}}$ ligne,
- $A_{\backslash l, \backslash m}$ la matrice A privée de sa $l^{\text{ème}}$ ligne et de sa $k^{\text{ème}}$ colonne.

Algorithme de mise à jour récursive :

- $C^{\text{tmp}} = C$
- $P^{\text{tmp}} = C^{-1}\mathbf{X}^T [\mathbf{1}, \boldsymbol{y}_1, ..., \boldsymbol{y}_\Lambda]$
- **for** $k = j + 1, ..., n$

1. $v_1 = C_{k,k-1}$
2. $v_2 = C_{k,k}$
3. **if** $|v_2| > |v_1|$ **then** swap v_2 and v_1 **endif**
4. Transformation de Givens pour trianguliser la matrice C dans laquelle la $j^{\text{ème}}$ ligne et la $j^{\text{ème}}$ colonne doivent être supprimées :
   — $w = \frac{v_2}{v_1}$
   — $q = \sqrt{1 + w^2}$
   — $c = \frac{\text{sign}(v_1)}{q}$
   — $s = w \times c$
   — $r = |v_1| \times q$
5. Transformation de Givens :
   — $C^{\text{tmp}}_{k,k-1} = r$
   — $C^{\text{tmp}}_{k,k} = 0$
   — **for** l in $[j, k+1, k+2, ..., n]$:
      — $w = C^{\text{tmp}}_{l,k-1} \times c + C^{\text{tmp}}_{l,k} \times s$
      — $C^{\text{tmp}}_{l,k} = -C^{\text{tmp}}_{l,k-1} \times s + C^{\text{tmp}}_{l,k} \times c$
      — $C^{\text{tmp}}_{l,k-1} = w$
   — **endfor**
6. Mise à jour du produit avec les données :
   — **for** i in $[1, ..., n]$:
      — $w = P^{\text{tmp}}_{k-1,i} \times c + P^{\text{tmp}}_{k,i} \times s$
      — $P^{\text{tmp}}_{k,i} = -P^{\text{tmp}}_{k-1,i} \times s + P^{\text{tmp}}_{k,i} \times c$
      — $P^{\text{tmp}}_{k-1,i} = w$
   — **endfor**

endfor

- $\boldsymbol{a} = P^{\text{tmp}}_j$

- Décalage des $n - j$ dernières lignes :
  for $k = j + 1, ..., n$:
  — $P^{\text{tmp}}_{k-1} = P^{\text{tmp}}_k$
  endfor

- La $j^{\text{ème}}$ est finalement placée dans la dernière ligne de P^{tmp} :
  $P^{\text{tmp}}_n = \boldsymbol{a}$

Cet algorithme permet, d'une part, d'effectuer les permutations et transformations sur la matrice C afin d'obtenir $\tilde{C}$ en supprimant la $j^{\text{ème}}$ ligne et la $n^{\text{ème}}$ colonne de C, et d'autre part, d'effectuer les permutations et transformations sur la matrice $C^{-1}\mathbf{X}^T[\mathbf{1}, \boldsymbol{y}_1, ..., \boldsymbol{y}_\Lambda]$ nécessaire à placer les éléments correspondant à l'objet à supprimer dans la dernière ligne notée $\boldsymbol{a}$:

$$P^{\text{tmp}} = \left(\begin{array}{c} \widetilde{C}^{-1}\widetilde{\mathbf{X}}^T[\mathbf{1}, \boldsymbol{y}_1, ..., \boldsymbol{y}_\Lambda] \\ \hline \boldsymbol{a} \end{array} \right) \tag{D.1}$$

L'algorithme décrit ci-dessus renvoie :
— $\boldsymbol{a} = P^{\text{tmp}}_n$, qui représente la dernière ligne de la matrice P^{tmp} et correspond à la contribution de l'objet à retirer de la configuration ;

— $\widetilde{C}^{-1}\widetilde{\mathbf{X}}^T[\mathbf{1}, \boldsymbol{y}_1, ..., \boldsymbol{y}_\Lambda] = P^{\text{tmp}}_{\backslash n}$, est la matrice P^{tmp} dont la dernière ligne a été supprimée ;

— $\widetilde{C} = C^{\text{tmp}}_{\backslash j,\backslash n}$, est la matrice C^{tmp} dont la $j^{\text{ème}}$ ligne et la $n^{\text{ème}}$ colonne ont été supprimées.

Finalement :

$$\begin{aligned}&[\mathbf{1}, \boldsymbol{y}_1,..., \boldsymbol{y}_\Lambda]^T\widetilde{\mathbf{X}}(\widetilde{C}\widetilde{C}^T)^{-1}\widetilde{\mathbf{X}}^T[\mathbf{1}, \boldsymbol{y}_1, ..., \boldsymbol{y}_\Lambda]\\ &\quad= [\mathbf{1}, \boldsymbol{y}_1, ..., \boldsymbol{y}_\Lambda]^T\mathbf{X}(CC^T)^{-1}\mathbf{X}^T[\mathbf{1}, \boldsymbol{y}_1, ..., \boldsymbol{y}_\Lambda] - \boldsymbol{a}^T\boldsymbol{a}.\end{aligned}$$

Si on note $g = C^{\text{tmp}}_{j,n}$ le dernière élément de la $j^{\text{ème}}$ ligne de C^{tmp}, le ratio des déterminants $\left(\frac{\det\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}}{\det\mathbf{X}^T\mathbf{X}}\right)$ présent dans le ratio de Métropolis-Hastings-Green pour une mort se simplifie par :

$$\left(\frac{\det\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}}{\det\mathbf{X}^T\mathbf{X}}\right) = \left(\frac{\det\widetilde{\mathbf{C}}\widetilde{\mathbf{C}}^T}{\det\mathbf{C}\mathbf{C}^T}\right) \tag{D.2}$$

$$= \frac{\left(\det\widetilde{\mathbf{C}}\right)^2}{\left(\det\mathbf{C}\right)^2} \tag{D.3}$$

$$= \frac{\left(\det C^{\text{tmp}}_{\backslash j,\backslash n}\right)^2}{\left(\det\mathbf{C}^{\text{tmp}}\right)^2} \tag{D.4}$$

$$= \left(\frac{\det C^{\text{tmp}}_{\backslash j,\backslash n}}{g \times \det C^{\text{tmp}}_{\backslash j,\backslash n}}\right)^2 \tag{D.5}$$

$$= \left(\frac{1}{g}\right)^2. \tag{D.6}$$

Le passage de l'équation (D.3) à l'équation (D.4) se justifie par le fait que les permutations effectuées sur C pour la transformer en C^{tmp} laissent le déterminant inchangé. De même, pour passage de l'équation (D.4) à l'équation (D.5), en décalant la $j^{\text{ème}}$ ligne de C^{tmp} à la $n^{\text{ème}}$ (et en faisant remonter les $n - j$ dernières lignes d'une ligne chacune), la matrice devient alors triangulaire inférieure et on a : $\det\mathbf{C}^{\text{tmp}} = g \times \det C^{\text{tmp}}_{\backslash j,\backslash n}$.

Tout ceci mène à l'expression du ratio de Metropolis-Hastings-Green donnée dans l'équation (2.30) :

$$r(\boldsymbol{u}, \boldsymbol{v}) = g^\Lambda \times \frac{1}{q} \times \prod_{\lambda=1}^{\Lambda} \exp\left\{\frac{-a^2_{\lambda+1} + 2m_\lambda a_1 a_{\lambda+1} - m^2_\lambda a^2_1}{2\sigma^2_\lambda}\right\} \times \frac{1}{\sqrt{2\pi\sigma^2_\lambda}}$$

Annexe E

Modélisation matricielle du filtrage adapté à la PSF en trois dimensions de l'instrument MUSE

Dans cette annexe, nous allons définir le filtrage adapté à la PSF de l'instrument MUSE sous forme matricielle. La variabilité spectrale de la FSF et de la LSF rend cette modélisation complexe, nous aurons besoin d'introduire un grand nombre de notations qui surchargeraient le corps du manuscrit. Dans le chapitre 3 où apparaît ce filtrage adapté, nous avons défini simplement le filtrage adapté sous forme matricielle de la façon suivante :

$$\mathbf{Y}^{f,(v)} = \mathbf{H}^T\mathbf{Y}$$

où le vecteur $\mathbf{Y}^{f,(v)}$ de taille $N \times 1$, avec $N = P \times Q \times \Lambda$, est la sortie du filtrage adapté sous forme vectorisée, la matrice $\mathbf{H}$ de taille $N \times N$ est la matrice contenant dans chaque colonne $h_{p,q,\lambda}$ la réponse de la PSF centrée en un point (p, q, λ) du cube sous forme vectorisée. Le vecteur $\mathbf{Y}$ de taille $N \times 1$ est le vecteur de données tel que $\mathbf{Y} = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_\lambda, \cdots, \boldsymbol{y}_\Lambda]^T$ avec $\boldsymbol{y}_\lambda$ un vecteur ligne de taille $(1 \times (P \times Q))$ représentant la vectorisation de l'image à la longueur d'onde λ du cube de données. Nous allons détailler dans cette annexe la construction de la matrice $\mathbf{H}$ à partir de l'équation de filtrage adapté (3.1).

E.1 Convolution par la FSF

L'équation (3.1) de filtrage adapté fait apparaître une convolution plan par plan (à μ fixé) entre la FSF et les données :

$$\Big(F_\mu * \mathbf{Y}(\cdot,\cdot,\mu)\Big)(p,q) = \sum_{z_p}\sum_{z_q} F_\mu(p - z_p, q - z_q)\mathbf{Y}(z_p, z_q, \mu)$$

Nous souhaitons traduire cette convolution sous forme matricielle : $\Big(F_\mu * \mathbf{Y}(\cdot,\cdot,\mu)\Big)(p,q) = F_{p,q,\mu}^{(vc)T}\boldsymbol{y}_\mu$, où $F_{p,q,\mu}^{(vc)}$ est le vecteur de taille $(P \times Q) \times 1$ qui correspond à la vectorisation de la FSF définie pour le point de coordonnées (p, q, μ) complétée par des zéros comme illustré sur la figure E.1. La vectorisation de la FSF complétée de zéros et de l'image à la longueur d'onde μ considérée, nous permet de traduit de façon équivalente la convolution en deux dimensions par un produit scalaire entre deux vecteurs $F_{p,q,\mu}^{(vc)}$ et $\boldsymbol{y}_\mu$.

Avec les notations introduites dans ce paragraphe, la modélisation schématique du cube de FSF défini pour la position spatiale (p, q) est donnée par la figure E.2.

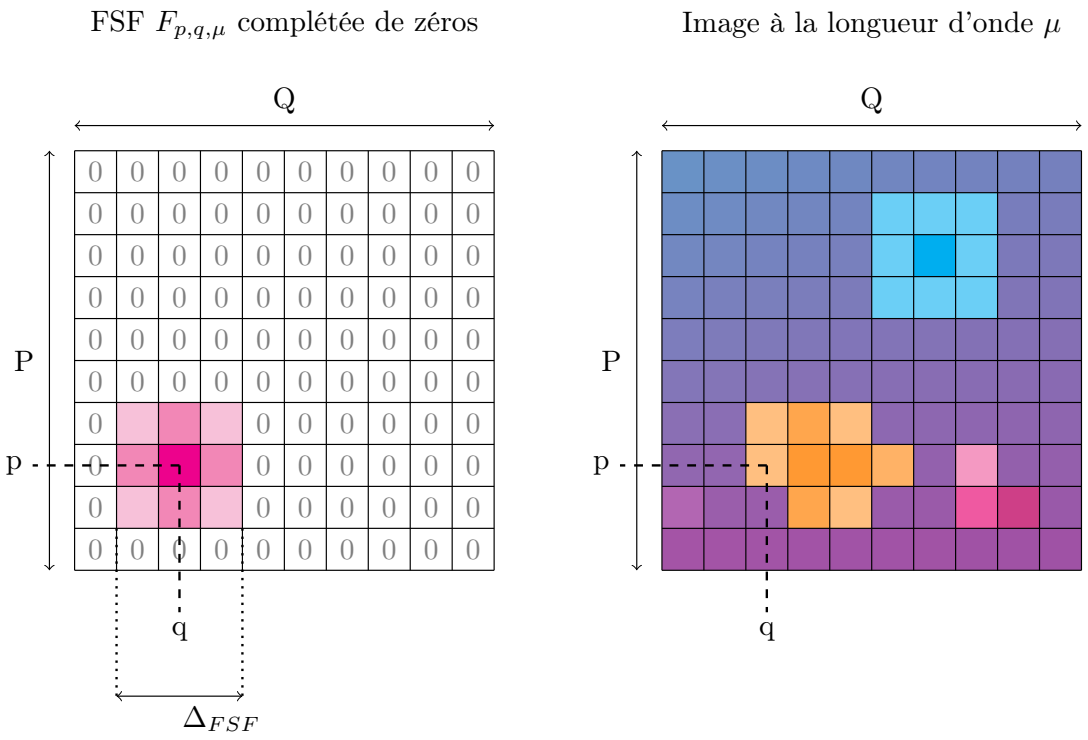


FIGURE E.1 – Représentation schématique de la FSF complétée de zéros (gauche) pour atteindre la taille d'une image du cube (droite). La FSF de dimension $\Delta_{FSF} \times \Delta_{FSF}$, avec $\Delta_{FSF} = 3$ pixels, est plongée dans une image de taille $P \times Q$, avec $P = Q = 10$ pixels dans cet exemple. Vectoriser la FSF complétée de zéros définie en (p, q, μ) revient à coller les colonnes de la FSF complétée de zéros les unes en dessous des autres pour former un vecteur $F^{(vc)}_{p,q,\mu}$ de taille $(P \times Q) \times 1$, la même opération est appliquée à l'image pour obtenir le vecteur $\boldsymbol{y}_\mu$.

E.2 Composition par la LSF

La composition par la LSF de l'instrument est la version discrète de l'intégrale de Fredholm qui se traduit par :

$$\begin{aligned}\mathbf{Y}^f(p,q,\lambda) &= \sum_{\mu=\lambda-\frac{\Delta_{LSF}}{2}}^{\lambda+\frac{\Delta_{LSF}}{2}} \widetilde{L}_\mu(\lambda)\Big\{\Big(F_\mu * \mathbf{Y}(\cdot,\cdot,\mu)\Big)(p,q)\Big\} \\ &= \sum_{\mu=\lambda-\frac{\Delta_{LSF}}{2}}^{\lambda+\frac{\Delta_{LSF}}{2}} \widetilde{L}_\mu(\lambda)\Big(F^{(vc)T}_{p,q,\mu}\boldsymbol{y}_\mu\Big),\end{aligned}$$

avec Δ_{LSF} la largeur de la LSF élargie. Cette opération se traduit par la multiplication du vecteur $F^{(vc)}_{p,q}$ de taille $N \times 1$ par une matrice diagonale par bloc $\mathbf{M}_{L,\lambda}$ de taille $N \times N$ où chaque bloc $(\mathbf{M}_{L,\lambda})_\mu$ avec $\mu = 1, \cdots, \Lambda$ de taille $(P \times Q) \times (P \times Q)$ est également diagonal. Les éléments de la diagonal d'un bloc sont identiques et valent $\widetilde{L}_\mu(\lambda)$ si $\big|\lambda - \mu\big| \leqslant \Delta_{LSF}/2$ et 0 sinon. La figure E.3 donne une représentation de cette matrice diagonale.

Le produit de cette matrice $\mathbf{M}_{L,\lambda}$ avec le vecteur $F^{(vc)}_{p,q}$, est le vecteur $h_{p,q,\lambda}$ représenté sur la figure E.4 qui correspond à la réponse du filtre adapté à la PSF spectralement élargie au point de coordonnées (p, q, λ) mais aussi à la réponse d'une source quasi-ponctuelle centrée en (p, q, λ).

$$
F_{p,q}^{(vc)} = \left(\begin{array}{c|c|c}
0 \times F_{p,q,1}^{(vc)} & \updownarrow P \times Q & \\
\vdots & \vdots & \\
\times F_{p,q,\lambda-4}^{(vc)} & \updownarrow P \times Q & \\
F_{p,q,\lambda-3}^{(vc)} & \updownarrow P \times Q & \\
F_{p,q,\lambda-2}^{(vc)} & \updownarrow P \times Q & \\
F_{p,q,\lambda-1}^{(vc)} & \updownarrow P \times Q & \\
F_{p,q,\lambda}^{(vc)} & \updownarrow P \times Q & P \times Q \times \Lambda \\
F_{p,q,\lambda+1}^{(vc)} & \updownarrow P \times Q & \\
F_{p,q,\lambda+2}^{(vc)} & \updownarrow P \times Q & \\
F_{p,q,\lambda+3}^{(vc)} & \updownarrow P \times Q & \\
\times F_{p,q,\lambda+4}^{(vc)} & \updownarrow P \times Q & \\
\vdots & \vdots & \\
0 \times F_{p,q,\Lambda}^{(vc)} & \updownarrow P \times Q &
\end{array} \right) \updownarrow
$$

FIGURE E.2 – Représentation schématique du cube de FSF défini pour la position spatiale (p, q) sous forme vectorisée et complétée de zéros.

Dans le chapitre 3 nous utiliserons directement la notation $h_{p,q,\lambda}$ sans revenir à la construction de ce vecteur.

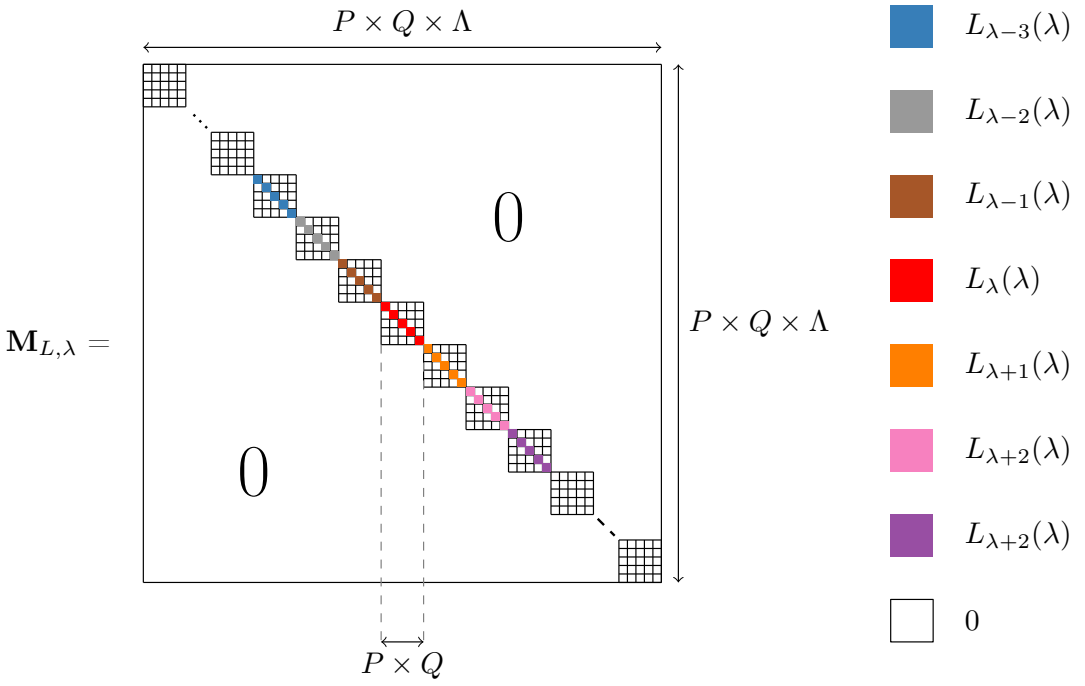


FIGURE E.3 – Représentation graphique de la matrice $\mathbf{M}_{L,\lambda}$, les élements diagonaux d'une même couleur symbolisent la valeur $\widetilde{L}_{\mu}(\lambda)$ constante dans la diagonale du μ^{eme} bloc. Dans cette représentation, la largeur de la LSF élargie est $\Delta_{LSF} = 7$ pixels.

E.3 Filtrage adapté à la PSF globale de l'instrument MUSE

Nous pouvons maintenant définir la matrice $\mathbf{H} = \mathbf{M}_L F_{p,q}^{(vc)}$ modélisant sous forme matricielle le filtrage adapté à la PSF, spectralement élargie, de MUSE pour l'intégralité du cube. La figure E.5 donne une représentation en fonction de la réponse du filtre $h_{p,q,\lambda}$ en chaque point de coordonnées (p, q, λ). Le filtrage adapté des données s'exprime alors comme le produit scalaire de la matrice de filtrage $\mathbf{H}$ et du vecteur de données $\mathbf{Y}$: $\mathbf{Y}^{f,(v)} = \mathbf{H}^T\mathbf{Y}$.

$$h_{p,q,\lambda} = \begin{pmatrix} 0 \times F^{(vc)}_{p,q,1} \\ \vdots \\ 0 \times \times F^{(vc)}_{p,q,\lambda-4} \\ L_{\lambda-3}(\lambda) \times F^{(vc)}_{p,q,\lambda-3} \\ L_{\lambda-2}(\lambda) \times F^{(vc)}_{p,q,\lambda-2} \\ L_{\lambda-1}(\lambda) \times F^{(vc)}_{p,q,\lambda-1} \\ L_{\lambda}(\lambda) \times F^{(vc)}_{p,q,\lambda} \\ L_{\lambda+1}(\lambda) \times F^{(vc)}_{p,q,\lambda+1} \\ L_{\lambda+2}(\lambda) \times F^{(vc)}_{p,q,\lambda+2} \\ L_{\lambda+3}(\lambda) \times F^{(vc)}_{p,q,\lambda+3} \\ 0 \times \times F^{(vc)}_{p,q,\lambda+4} \\ \vdots \\ 0 \times F^{(vc)}_{p,q,\Lambda} \end{pmatrix} \begin{matrix} \updownarrow P \times Q \\ \vdots \\ \updownarrow P \times Q \\ \updownarrow P \times Q \\ \updownarrow P \times Q \\ \updownarrow P \times Q \\ \updownarrow P \times Q \\ \updownarrow P \times Q \\ \updownarrow P \times Q \\ \updownarrow P \times Q \\ \updownarrow P \times Q \\ \vdots \\ \updownarrow P \times Q \end{matrix} \quad P \times Q \times \Lambda$$

FIGURE E.4 – Représentation matricielle de la PSF de l'instrument MUSE pour le filtrage adapté. Dans cette représentation, la largeur de la LSF élargie est $\Delta_{LSF} = 7$ pixels.

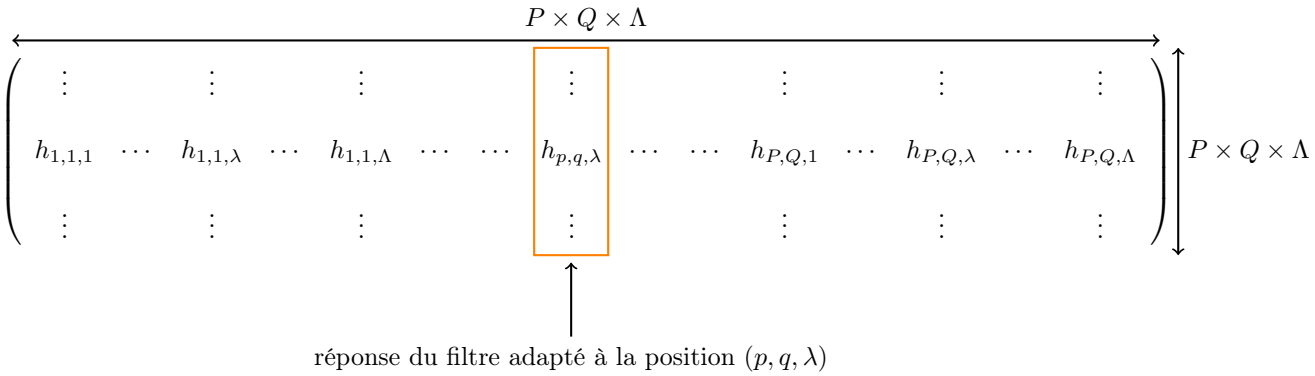


FIGURE E.5 – Modélisation matricielle de l'opération de filtrage adapté.

Annexe F

Détail de l'algorithme du sigma-clipping par point fixe

Soit $\boldsymbol{y} = [y_1, \cdots, y_N]$ les données d'une image du cube de données sous forme vectorisée. Pour simplifier les notations, l'indiçage par λ sera omis dans cette annexe. Notons F la fonction de répartition empirique des données. Cette fonction de répartition tient compte de la distribution des données sous l'hypothèse $\mathcal{H}_0$, où il n'y a que du bruit et de la distribution des données sous l'hypothèse $\mathcal{H}_1$. Le mélange de lois s'écrit $F = \pi_0 F_0 + \pi_1 F_1$, où π_0 est la proportion de données sous $\mathcal{H}_0$ et F_0, la fonction de répartition des données sous $\mathcal{H}_0$, et de même pour π_1 et F_1 qui sont la proportion de données et la fonction de répartition sous $\mathcal{H}_1$. Afin de réaliser l'estimation de la médiane et de la variance des données sous la loi $\mathcal{H}_0$, une opération de troncature (*clipping*) est appliquée aux données afin d'éliminer en partie les données contaminées par les sources. La fonction de répartition empirique des données tronquées F_t peut s'exprimer en fonction de la répartition F :

$$F_t(y) = \frac{F(y) - F(g)}{F(d) - F(g)} \tag{F.1}$$

où g est le seuil de troncature à gauche et d le seuil de troncature à droite. Sous l'hypothèse que F_0 est symétrique, les seuils sont alors choisis comme :

$$g = \mu - \kappa\sigma$$
$$d = \mu + \kappa\sigma$$

où (μ, σ) désignent la médiane et l'écart-type de la loi tronquée F_t (qui doivent être similaires à ceux sur F_0 si le clipping est suffisamment strict), et κ est un paramètre d'échelle qui définit l'importance de la troncature effectuée. Les quantités nécessaires aux calculs de la médiane et de l'écart-type sont illustrées sur la figure F.1.

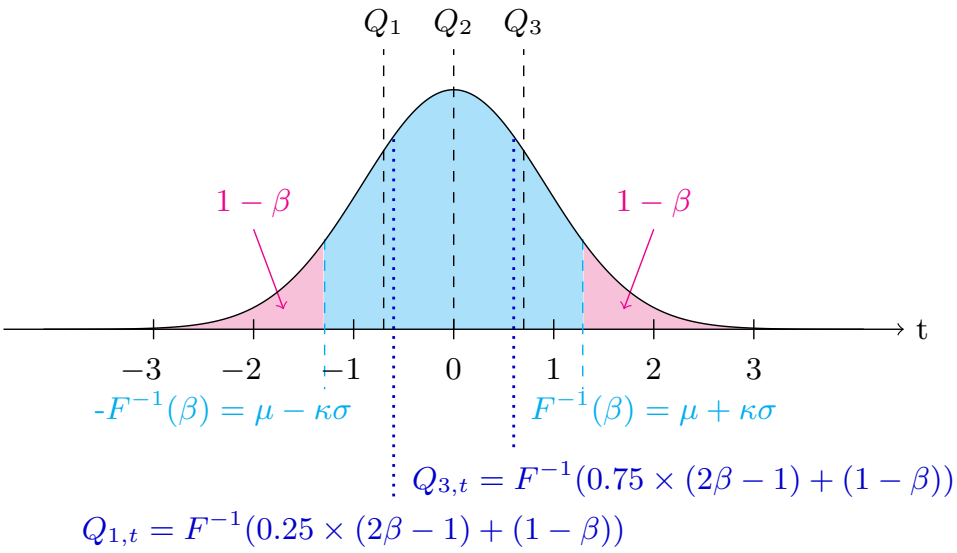


FIGURE F.1 – Représentation de la loi des données et des différentes quantités nécessaires à la réduction du cube de données : les quartiles $Q1$, $Q2$, et $Q3$ à 25%, à 50% (médiane = moyenne) et à 75%, le paramètre β de troncature des données pour l'estimation de la médiane et de l'écart-type.

F.1 Estimation de la médiane

La médiane μ des données tronquées est telle que $F_t(\mu) = \frac{1}{2}$ ce qui permet d'écrire :

$$\begin{aligned} F_t(\mu) = \frac{F(\mu) - F(g)}{F(d) - F(g)} &\iff \mu = F_t^{-1}\left(\frac{F(\mu) - F(g)}{F(d) - F(g)}\right) \\ &\iff F_t^{-1}\left(\frac{1}{2}\right) = F_t^{-1}\left(\frac{F(\mu) - F(g)}{F(d) - F(g)}\right) \\ &\iff \frac{1}{2} = \frac{F(\mu) - F(g)}{F(d) - F(g)} \\ &\iff F(\mu) = F(g) + \frac{1}{2}\left(F(d) - F(g)\right) \\ &\iff \mu = F^{-1}\left(\frac{F(d) + F(g)}{2}\right) \end{aligned} \tag{F.2}$$

F.2 Estimation de l'écart-type

L'estimation de l'écart-type est paramétrique puisque nous utilisons un modèle gaussien tronqué pour modéliser les données tronquées, l'expression de l'estimateur de l'écart type est donnée

par :

$$\sigma = \frac{Q_{3,t} - Q_{1,t}}{\mathcal{F}} \iff \sigma = \frac{2(\mu - Q_{1,t})}{2\mathcal{F}_1} \tag{F.3}$$

où $Q_{1,t}$ et $Q_{3,t}$ sont les premier et troisième quartile des données tronquées. $\mathcal{F}$ est le facteur correctif permettant de prendre en compte la troncature de la loi gaussienne centrée réduite. L'équation (F.3) revient à identifier l'écart interquartile $Q_{3,t} - Q_{1,t}$ des données tronquées avec l'écart interquartile $\sigma\mathcal{F} = \sigma\left(Q^{\mathcal{N}}_{3,t} - Q^{\mathcal{N}}_{1,t}\right) = \sigma\left(2(Q^{\mathcal{N}}_{2,t} - Q^{\mathcal{N}}_{1,t})\right) = \sigma\left(2\mathcal{F}_1\right)$ de la loi gaussienne de moyenne nulle et d'écart-type σ que nous utilisons pour modéliser les données.

Le premier quartile des données tronquées, $Q_{1,t}$ est tel que $F_t(Q_{1,t}) = \frac{1}{4}$, et la fonction de répartition F_t des données tronquées peut s'exprimer en fonction de F :

$$\begin{aligned} F_t(Q_{1,t}) = \frac{F(Q_{1,t}) - F(g)}{F(d) - F(g)} &\iff \frac{F(Q_{1,t}) - F(g)}{F(d) - F(g)} = \frac{1}{4} \\ &\iff F(Q_{1,t}) = \frac{1}{4}(F(d) - F(g)) + F(g) \\ &\iff Q_{1,t} = F^{-1}\left(\frac{F(d) + 3F(g)}{4}\right) \end{aligned} \tag{F.4}$$

Le facteur correctif $\mathcal{F}_1$ peut s'écrire :

$$\mathcal{F}_1 = \Phi^{-1}\left(\frac{1}{2}(\Phi(\kappa) - \Phi(-\kappa)) + \Phi(-\kappa)\right) - \Phi^{-1}\left(\frac{1}{4}(\Phi(\kappa) - \Phi(-\kappa)) + \Phi(-\kappa)\right)$$

avec Φ la fonction de répartition de la loi normale. Or Φ est symétrique donc :

$$\begin{aligned} \Phi^{-1}\left(\frac{1}{2}(\Phi(\kappa) - \Phi(-\kappa)) + \Phi(-\kappa)\right) &= \Phi^{-1}\left(\frac{1}{2}(\Phi(\kappa) + \Phi(-\kappa))\right) \\ &= \Phi^{-1}\left(\frac{1}{2}\right) \\ &= 0 \end{aligned} \tag{F.5}$$

et

$$\begin{aligned} -\Phi^{-1}\left(\frac{1}{4}(\Phi(\kappa) - \Phi(-\kappa)) + \Phi(-\kappa)\right) &= -\Phi^{-1}\left(\frac{1}{4}(\Phi(\kappa) + 3\Phi(-\kappa))\right) \\ &= \Phi^{-1}\left(1 - \frac{1}{4}\Big(\Phi(\kappa) + 3\Phi(-\kappa)\Big)\right) \\ &= \Phi^{-1}\left(1 - \frac{1}{4}\Big(\Phi(\kappa) + 3\big(1 - \Phi(\kappa)\big)\Big)\right) \\ &= \Phi^{-1}\left(\frac{1}{2}\left(\frac{1}{2} + \Phi(\kappa)\right)\right) \end{aligned} \tag{F.6}$$

Le facteur correctif s'écrit donc $\mathcal{F}_1 = \Phi^{-1}\left(\frac{1}{2}\left(\frac{1}{2} + \Phi(\kappa)\right)\right)$.

F.3 Algorithme du point fixe

Finalement, les estimateurs de médiane μ et d'écart-type σ utilisée dans l'encadré 3.5. sont obtenus, dans le cas continu, à l'aide d'un algorithme du point fixe où les expressions :

$$\begin{aligned} \mu &= F^{-1}\left(\frac{F(\mu + \kappa\sigma) + F(\mu - \kappa\sigma)}{2}\right) \\ \sigma &= \frac{1}{\mathcal{F}_1}\left(\mu - F^{-1}\left(\frac{F(\mu + \kappa\sigma) + 3F(\mu - \kappa\sigma)}{4}\right)\right) \end{aligned}$$

des estimateurs peuvent se résumer par $(\mu, \sigma) = g(\mu, \sigma)$. Pour montrer que l'algorithme du point fixe converge, il faut montrer que la fonction g est contractante, en montrant par exemple que les valeurs propres de la jacobienne de g sont, en valeur absolue, inférieures à 1 au moins dans le voisinage du point fixe. Notons qu'expérimentalement, la fonction g s'avère toujours très plate autour de la valeur du point fixe, ce qui explique la convergence. La troncature des données est réalisée à chaque itération k à l'aide des valeurs estimées de la médiane μ_{k-1} et de l'écart-type σ_{k-1} :

$$\boldsymbol{y}_t^{(k)} = \Big\{ \boldsymbol{y}_i, i = 1, \cdots, N \text{ t.q. } \Big| \boldsymbol{y}_i - \mu_{k-1} \Big| \leqslant \kappa \sigma_{k-1} \Big\}.$$

Contrairement aux algorithmes itératifs classiques de σ-clipping, où les données sont tronquées de manière récursives, les données sont ici tronquées une seule fois : seul l'intervalle de troncature diffère d'une itération à l'autre.

En pratique, la fonction F est remplacée dans l'algorithme par la fonction de répartition empirique $\bar{F}$ des données qui est une fonction en escalier (valeurs discrètes de la fonction de répartition empirique), et F^{-1} est remplacée par l'inverse généralisée de $\bar{F}$ par exemple. L'avantage du cas discret est que la convergence de l'algorithme itératif par point fixe est garantie en un nombre fini d'itérations ; chaque itération permet de diminuer la distance au point fixe, et il existe un nombre fini d'intervalles de troncature équivalents. En revanche la précision de l'estimation est limitée par le nombre d'échantillons dans les données.

Annexe G

Détail de la procédure de contrôle du FDR par le knockoff filter

Barber and Candes [2014] se sont intéressés au problème du contrôle du FDR dans les procédures de sélection de variables pour un modèle de mélange linéaire gaussien :

$$\boldsymbol{y} = \mathbf{X}\beta + \boldsymbol{z},$$

où $\boldsymbol{y} \in \mathbb{R}^n$ est le vecteur des observations, $\mathbf{X} = [\mathbf{X}_1, \cdots, \mathbf{X}_p] \in \mathbb{R}^{n \times p}$ est une matrice de design connue, $\beta \in \mathbb{R}^p$ est le vecteur de coefficients à estimer et $\boldsymbol{z} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$ est un bruit gaussien. Pour garantir l'existence d'une unique solution, nous nous intéresserons aux cas où $n \geqslant p$. C'est en pratique souvent le cas, notamment avec l'apparition des données massives, nous disposons d'un très grand nombre d'observations qui peuvent s'expliquer avec un petit nombre de variables, ce qui veut dire aussi que le nombre de coefficients non nuls de β est supposé petit.

Barber and Candes [2014] montrent qu'il est possible de contrôler le FDR pour n'importe quelle procédure de sélection qui retourne un sous ensemble $\hat{S} \subset \{1, 2, \cdots, p\}$ de variables. Le FDR s'écrit alors :

$$\text{FDR} = \text{E}\left[\frac{\#\{j : \beta_j = 0, j \in \hat{S}\}}{\max(\#\{j : j \in \hat{S}\}, 1)}\right] \tag{G.1}$$

Le principe du *knockoff filter* consiste à construire des contrefaçons (*knockoff*) des variables stockées dans la matrice $\mathbf{X}$. Ces contrefaçons doivent respecter certaines contraintes que nous détaillerons par la suite et le principe de la méthode repose sur l'idée que si les variables $\mathbf{X}_j$ ne contribuent pas aux observations, il y a la même probabilité de sélectionner la variable ou sa contrefaçon avec la procédure de sélection choisie. Le *knockoff filter* se décompose en plusieurs étapes :

Construction des contrefaçons : soit $\Sigma = \mathbf{X}^T\mathbf{X}$ la matrice de Gram des variables $\mathbf{X}_j$ originales, normalisées telles que $\Sigma_{jj} = \|\mathbf{X}_j\|_2^2 = 1$ pour tout j. Pour chaque composante $\mathbf{X}_j$, construisons une contrefaçon $\tilde{\mathbf{X}}_j$ telle que :

$$\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} = \Sigma \tag{G.2}$$

$$\mathbf{X}^T\tilde{\mathbf{X}} = \Sigma - \text{diag}\{\boldsymbol{s}\} \tag{G.3}$$

$$\mathbf{X}_j^T\tilde{\mathbf{X}}_k = \mathbf{X}_j^T\mathbf{X}_k \text{ , pour tout } j \neq k \tag{G.4}$$

Les contrefaçons ont la même structure de corrélations que les variables originales, équations (G.2) et (G.4), en revanche la variable et sa contrefaçon sont les moins corrélées possible, l'équation (G.3) se traduit par $\mathbf{X}_j^T\tilde{\mathbf{X}}_j = \Sigma_{jj} - s_j$. Pour construire de bonnes contrefaçons, il faut choisir les s_j aussi grands que possible. Une des stratégies proposées

par les auteurs est de choisir $\boldsymbol{s} \in \mathbb{R}_+^p$ tel que $\text{diag}(\boldsymbol{s}) \leqslant 2\Sigma$ et construire la matrice $\tilde{\mathbf{X}}$ telle que :

$$\tilde{\mathbf{X}} = \mathbf{X}\left(\mathbf{I} - \Sigma^{-1}\text{diag}(\boldsymbol{s})\right) + \tilde{\mathbf{U}}\mathbf{C},$$

avec $\tilde{\mathbf{U}}$ une matrice de taille $n \times p$ orthonormale, orthogonale au sous-espace des variables de $\mathbf{X}$, et $\mathbf{C}$ est la décomposition de Cholesky : $\mathbf{C}^T\mathbf{C} = 2\text{diag}\{\boldsymbol{s}\} - \text{diag}\{\boldsymbol{s}\}\Sigma^{-1}\text{diag}\{\boldsymbol{s}\}$.

Calcul de la statistique de test : il s'agit ici de construire une statistique W_j de test pour chaque paire $(\mathbf{X}_j, \tilde{\mathbf{X}}_j)$ de variables et leurs contrefaçons afin de décider quelles sont les variables qui contribuent aux observations. Les auteurs donnent différentes approches pour construire cette statistique de test W. Ce qu'il faut retenir c'est que W doit dépendre de $\mathbf{X}$, $\tilde{\mathbf{X}}$ et $\boldsymbol{y}$, et prend une valeur W_j positive significative lorsqu'il est fort probable que $\beta_j \neq 0$, *i.e.* que la variable $\mathbf{X}_j$ contribue aux observations et de manière beaucoup plus convaincante que sa contrefaçon.

Calcul du seuil T à partir des données : La dernière étape du *knockoff filter* consiste à calculer le seuil à l'aide duquel nous allons ensuite seuiller les W_j et donc les β_j correspondant. Si q désigne le taux de fausses découvertes que nous souhaitons garantir lors de la phase de sélection de variables, le seuil T s'écrit :

$$T = \min\left\{t \in \mathcal{W} : \frac{\#\{j : W_j \leqslant -t\}}{\max\left(\#\{j : W_j \geqslant\}, 1\right)} \leqslant q\right\},$$

où $\mathcal{W} = \{|W_j| : j = 1, \cdots, p\}\backslash\{0\}$. La fraction qui apparait dans l'expression du seuil est une estimation du FDR.

Finalement la sélection de variable est effectuée de la façon suivante :

$$\hat{S} = \{j : W_j \geqslant T\},$$

avec T le seuil calculé à partir des données pour un niveau de contrôle q fixé. D'après le théorème 1 énoncé dans Barber and Candes [2014], pour tout $q \in [0, 1]$, le *knockoff filter* garantit :

$$\text{E}\left[\frac{\#\{j : \beta_j = 0, j \in \hat{S}\}}{\#\{j : j \in \hat{S}\} + q^{-1}}\right] \leqslant q.$$

Cette expression est très proche du contrôle du taux de fausses découvertes défini par l'équation (G.1). Pour un contrôle exact du FDR, les auteurs définissent une alternative (*knockoff+*) qui entraîne une modification mineure du seuil :

$$T = \min\left\{t \in \mathcal{W} : \frac{1 + \#\{j : W_j \leqslant -t\}}{\max\left(\#\{j : W_j \geqslant\}, 1\right)} \leqslant q\right\},$$

qui produit un seuillage un peu plus conservatif.

Annexe H

Résultats de la détection de galaxies sur le cube HDFS

Dans cette annexe, nous donnons la liste des objets détectés dans le cas présenté dans le chapitre 3 qui ne sont pas répertoriés dans le catalogue HST.

ID	RA	DEC	axe 1	axe 2	angle	indice Sersic	étape	λ_{max}
6	338.24812941	-60.5593519826	2.54	4.59	0.40	2.0	0	-1.0
13	338.248089423	-60.5551180395	5.73	5.39	1.52	1.0	0	-1.0
60	338.232912329	-60.5550824936	4.44	4.67	0.01	2.0	0	-1.0
78	338.239579021	-60.5591966058	3.12	1.92	1.03	0.5	0	-1.0
82	338.235255196	-60.5679258656	2.22	2.61	0.36	0.5	0	-1.0
96	338.238930039	-60.5691439657	3.00	2.91	0.48	2.0	0	-1.0
108	338.240088491	-60.5679169678	1.09	1.95	0.80	0.5	0	-1.0
117	338.214567977	-60.5551437644	3.63	2.84	0.25	1.0	0	-1.0
125	338.223266599	-60.5551594822	2.73	1.72	0.32	0.5	0	-1.0
127	338.214790002	-60.5558079807	5.70	3.65	0.87	2.0	1	8340.0
130	338.228707883	-60.5699477857	2.89	5.61	1.20	2.0	1	6783.75
134	338.214764566	-60.5565290673	5.40	5.56	1.21	2.0	1	8351.25
139	338.248318238	-60.5583297757	2.92	5.58	1.45	2.0	1	8616.25
140	338.226128263	-60.559774312	4.03	2.86	0.52	2.0	1	6610.0
144	338.238276478	-60.5673049354	5.34	4.25	0.83	2.0	1	7548.75
152	338.238936286	-60.5592625948	3.23	2.40	0.61	0.5	1	7616.25
153	338.214642147	-60.5561858303	4.04	3.94	1.05	2.0	1	8282.5
159	338.247485407	-60.5552010516	3.40	2.80	1.24	2.0	1	7746.25
161	338.216608436	-60.5556596014	4.39	4.00	0.26	2.0	1	5230.0
163	338.213965061	-60.5575374496	3.22	3.07	1.10	0.5	1	8282.5
166	338.232794409	-60.5631036739	3.10	4.70	0.55	2.0	1	7173.75
171	338.218793654	-60.5716277132	3.24	1.70	0.98	0.5	1	7708.75
174	338.21406549	-60.5588143356	2.44	4.84	0.11	2.0	1	8288.75
176	338.243281254	-60.5549877019	5.01	5.99	0.23	2.0	1	8340.0
182	338.24239162	-60.5549220291	2.10	2.27	0.43	2.0	1	8833.75
183	338.247839958	-60.5556972655	4.88	5.94	0.36	2.0	1	8395.0
184	338.238995275	-60.5556134318	5.10	5.84	1.02	2.0	1	5022.5
189	338.214087052	-60.5580785591	5.90	3.87	0.72	2.0	1	8761.25
190	338.244720631	-60.5634689392	3.24	2.92	0.31	2.0	1	7173.75
192	338.214309527	-60.5585536197	3.40	3.45	1.24	2.0	1	8767.5

198	338.242722515	-60.560844892	4.20	3.57	0.28	2.0	1	6778.75
201	338.214176773	-60.555914654	2.62	2.073	0.09	2.0	1	8338.75
209	338.214599258	-60.5573921954	3.32	1.96	1.18	0.5	1	7988.75
211	338.214746369	-60.5576732454	3.28	3.035	0.38	0.5	1	8833.75
212	338.247878442	-60.5694984618	3.03	3.13	1.32	0.5	1	8340.0
218	338.239384683	-60.5674665066	2.33	4.14	1.20	2.0	1	7490.0
221	338.247602155	-60.555429567	3.19	2.18	0.64	0.5	1	8338.75
228	338.214945606	-60.5569768303	3.30	3.26	1.07	0.5	1	8340.0
229	338.214736852	-60.5590919608	5.50	5.71	0.61	2.0	1	8283.75
230	338.214370817	-60.5654001475	5.29	5.45	0.94	2.0	1	8351.25
236	338.218831709	-60.5713968949	2.77	1.49	0.95	0.5	1	7272.5
239	338.247725686	-60.5715406464	2.46	4.11	1.14	2.0	1	8833.75
241	338.215189027	-60.5549784013	5.70	4.071	0.46	2.0	1	8340.0
243	338.242051756	-60.5675091305	5.55	5.81	0.60	2.0	1	8011.25
245	338.214449788	-60.5719347797	3.24	3.11	0.91	0.5	1	8340.0
259	338.247344178	-60.5695256877	3.22	2.84	0.45	0.5	1	8021.25
266	338.214492631	-60.5553717992	3.26	1.97	0.43	0.5	1	8340.0
267	338.246896756	-60.5549948968	3.29	2.44	0.45	0.5	1	8338.75
272	338.239514073	-60.5677910257	3.08	3.18	1.08	0.5	1	7425.0
276	338.244094564	-60.5548369582	2.23	2.97	1.52	0.5	1	8298.75
278	338.214152956	-60.5573023434	2.84	2.75	0.96	0.5	1	7988.75
284	338.236890288	-60.5555118543	3.13	3.14	0.75	0.5	1	5330.0
293	338.21896742	-60.5711090643	3.29	2.75	1.35	0.5	1	8340.0
294	338.214555156	-60.5672733962	3.27	3.13	0.31	0.5	1	8282.5

Tableau H.1 – Liste des objets détectés dans le cas présenté dans le chapitre 3 qui ne sont pas répertoriés dans le catalogue HST. Les positions sont données en coordonnées du ciel (RA, DEC) qui permettent de préserver la position exacte des sources détectées quelle que soit la troncature du cube utilisée. La taille des axes est données en pixels, l'orientation est un angle en radian. L'indice Sersic permet de décrire la décroissance plus ou moins accentuée du profil d'intensité spatial de la galaxie. L'indice de détection vaut 0 si l'objet a été détecté sur l'image blanche, et 1 s'il a été détecté sur le cube complet. La longueur d'onde λ_{max} est donnée en Angstrom.

Annexe I

Marginalisation de la densité *a posteriori* jointe des paramètres de la configuration d'objet et du bruit

Dans cette annexe, nous allons présenter le détail des calculs de marginalisation de la densité *a posteriori*, définie à l'équation (2.24) et qui mène à l'expression définie dans l'équation (2.25). Nous présenterons en complément la marginalisation de la densité (2.25) qui permet d'obtenir les lois *a posteriori* conditionnelles des paramètres $\boldsymbol{m}$ et $\boldsymbol{\sigma^2}$.

I.1 Marginalisation par rapport aux intensités $\boldsymbol{w}_\lambda$

La marginalisation par rapport au paramètre d'intensité $\boldsymbol{w}_\lambda$ nécessite de développer le terme encadré en bleu dans l'équation suivante :

$$p(\boldsymbol{u}, \mathbf{W}, \boldsymbol{m}, \boldsymbol{\sigma^2}|\mathbf{Y}) \propto \prod_{\lambda=1}^{\Lambda} \Big\{ p(\boldsymbol{w}_\lambda, m_\lambda, \sigma_\lambda^2|\boldsymbol{u}, \boldsymbol{y}_\lambda) \Big\} \, p(\boldsymbol{u}) \tag{I.1}$$

Ce terme correspond à la densité *a posteriori* jointe de $\boldsymbol{w}_\lambda, m_\lambda, \sigma_\lambda^2$ conditionnellement à une configuration $\boldsymbol{u}$ d'objets données à la longueur d'onde λ fixée.

Afin de marginaliser la densité $p(\boldsymbol{w}_\lambda, m_\lambda, \sigma_\lambda^2|\boldsymbol{u}, \boldsymbol{y}_\lambda)$, il faut isoler tous les termes où $\boldsymbol{w}_\lambda$ intervient. Nous allons développer le terme encadré en rose dans l'équation suivante :

$$\begin{aligned} p(\boldsymbol{w}_\lambda, m_\lambda, \sigma_\lambda^2|\boldsymbol{u}, \boldsymbol{y}_\lambda) = & \left(\frac{1}{2\pi\sigma_\lambda^2}\right)^{\frac{M}{2}} \exp\left(-\frac{(\boldsymbol{y}_\lambda - \mathbf{X}_\lambda \boldsymbol{w}_\lambda - \mathbf{1}m_\lambda)^T(\boldsymbol{y}_\lambda - \mathbf{X}_\lambda \boldsymbol{w}_\lambda - \mathbf{1}m_\lambda)}{2\sigma_\lambda^2}\right) \\ & \times \frac{1}{\sigma_\lambda^2} \mathbb{1}_{]0,+\infty[}(\sigma_\lambda^2) \mathbb{1}_{\mathbb{R}^{n(\boldsymbol{u})}}(\boldsymbol{w}_\lambda) \end{aligned} \tag{I.2}$$

Notons $A = (\boldsymbol{y}_\lambda - \mathbf{X}_\lambda \boldsymbol{w}_\lambda - \mathbf{1}m_\lambda)^T(\boldsymbol{y}_\lambda - \mathbf{X}_\lambda \boldsymbol{w}_\lambda - \mathbf{1}m_\lambda)$ et développons A :

$$A = (\boldsymbol{y}_\lambda - \mathbf{X}_\lambda \boldsymbol{w}_\lambda - \mathbf{1}m_\lambda)^T(\boldsymbol{y}_\lambda - \mathbf{X}_\lambda \boldsymbol{w}_\lambda - \mathbf{1}m_\lambda) \tag{I.3}$$

$$= (\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda)^T(\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda) - 2\boldsymbol{w}_\lambda^T \mathbf{X}_\lambda^T(\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda) + \boldsymbol{w}_\lambda^T \mathbf{X}_\lambda^T \mathbf{X} \boldsymbol{w}_\lambda \tag{I.4}$$

On reconnait dans l'équation (I.4) le début du développement de l'identité remarquable $(\boldsymbol{w}_\lambda - \mu_\lambda)^T(\mathbf{X}_\lambda^T \mathbf{X}_\lambda)(\boldsymbol{w}_\lambda - \mu_\lambda)$ en posant :

$$\mu_\lambda = (\mathbf{X}_\lambda^T \mathbf{X}_\lambda)^{-1} \mathbf{X}_\lambda (\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda)$$

On a alors :

$$\begin{aligned}
A &= (\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda)^T(\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda) + (\boldsymbol{w}_\lambda - \mu_\lambda)^T(\mathbf{X}_\lambda^T\mathbf{X}_\lambda)(\boldsymbol{w}_\lambda - \mu_\lambda) - \mu_\lambda^T(\mathbf{X}_\lambda^T\mathbf{X}_\lambda)\mu_\lambda \\
&= (\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda)^T(\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda) + (\boldsymbol{w}_\lambda - \mu_\lambda)^T(\mathbf{X}_\lambda^T\mathbf{X}_\lambda)(\boldsymbol{w}_\lambda - \mu_\lambda) \\
&\qquad - (\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda)^T\mathbf{X}_\lambda\cancel{(\mathbf{X}_\lambda^T\mathbf{X}_\lambda)^{-1}}\cancel{(\mathbf{X}_\lambda^T\mathbf{X}_\lambda)}(\mathbf{X}_\lambda^T\mathbf{X}_\lambda)^{-1}\mathbf{X}_\lambda^T(\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda) \\
&= (\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda)^T(\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda) + (\boldsymbol{w}_\lambda - \mu_\lambda)^T(\mathbf{X}_\lambda^T\mathbf{X}_\lambda)(\boldsymbol{w}_\lambda - \mu_\lambda) \\
&\qquad - (\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda)^T\mathbf{X}_\lambda(\mathbf{X}_\lambda^T\mathbf{X}_\lambda)^{-1}\mathbf{X}_\lambda^T(\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda) \\
&= (\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda)^T(\mathbf{I}_M - \mathbf{X}_\lambda(\mathbf{X}_\lambda^T\mathbf{X}_\lambda)^{-1}\mathbf{X}_\lambda^T)(\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda) + (\boldsymbol{w}_\lambda - \mu_\lambda)^T(\mathbf{X}_\lambda^T\mathbf{X}_\lambda)(\boldsymbol{w}_\lambda - \mu_\lambda)
\end{aligned} \tag{I.5}$$

Finalement, la densité $p(\boldsymbol{w}_\lambda, m_\lambda, \sigma_\lambda^2|\boldsymbol{u}, \boldsymbol{y}_\lambda)$ peut se réécrire :

$$\begin{aligned}
p(\boldsymbol{w}_\lambda, m_\lambda, \sigma_\lambda^2|\boldsymbol{u}, \boldsymbol{y}_\lambda) = &\left(\frac{1}{2\pi\sigma_\lambda^2}\right)^{\frac{M}{2}} \frac{1}{\sigma_\lambda^2}\mathbb{1}_{]0,+\infty[}(\sigma_\lambda^2) \\
&\times \exp\left(-\frac{(\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda)^T(\mathbf{I}_M - \mathbf{X}_\lambda(\mathbf{X}_\lambda^T\mathbf{X}_\lambda)^{-1}\mathbf{X}_\lambda^T)(\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda)}{2\sigma_\lambda^2}\right) \\
&\times \exp\left(-\frac{(\boldsymbol{w}_\lambda - \mu_\lambda)^T(\mathbf{X}_\lambda^T\mathbf{X}_\lambda)(\boldsymbol{w}_\lambda - \mu_\lambda)}{2\sigma_\lambda^2}\right)
\end{aligned} \tag{I.6}$$

et l'opération de marginalisation se déduit de la façon suivante :

$$\begin{aligned}
\int p(\boldsymbol{w}_\lambda, m_\lambda, \sigma_\lambda^2|\boldsymbol{u}, \boldsymbol{y}_\lambda)d\boldsymbol{w}_\lambda = &\\
&\left(\frac{1}{2\pi\sigma_\lambda^2}\right)^{\frac{M}{2}-\frac{n(\boldsymbol{u})}{2}} \left|\left(\mathbf{X}_\lambda^T\mathbf{X}_\lambda\right)^{-1}\right|^{\frac{1}{2}} \frac{1}{\sigma_\lambda^2}\mathbb{1}_{]0,+\infty[}(\sigma_\lambda^2) \\
&\quad\times \exp\left(-\frac{(\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda)^T(\mathbf{I}_M - \mathbf{X}_\lambda(\mathbf{X}_\lambda^T\mathbf{X}_\lambda)^{-1}\mathbf{X}_\lambda^T)(\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda)}{2\sigma_\lambda^2}\right) \\
&\quad\times \int \left(\frac{1}{2\pi\sigma_\lambda^2}\right)^{\frac{n(\boldsymbol{u})}{2}} \frac{1}{\left|\left(\mathbf{X}_\lambda^T\mathbf{X}_\lambda\right)^{-1}\right|^{\frac{1}{2}}} \exp\left(-\frac{(\boldsymbol{w}_\lambda - \mu_\lambda)^T(\mathbf{X}_\lambda^T\mathbf{X}_\lambda)(\boldsymbol{w}_\lambda - \mu_\lambda)}{2\sigma_\lambda^2}\right)d\boldsymbol{w}_\lambda
\end{aligned} \tag{I.7}$$

L'intégrale de la loi de probabilité $\boldsymbol{w}_\lambda \sim \mathcal{N}\left(\mu_\lambda, (\mathbf{X}_\lambda^T\mathbf{X}_\lambda)^{-1}\right)$ vaut 1, ce qui permet d'écrire la densité *a posteriori* décrite dans l'équation (2.25) du manuscrit, pour une longueur d'onde λ donnée :

$$\begin{aligned}
\int p(\boldsymbol{w}_\lambda, m_\lambda, \sigma_\lambda^2|\boldsymbol{u}, \boldsymbol{y}_\lambda)d\boldsymbol{w}_\lambda = &\left(\frac{1}{2\pi\sigma_\lambda^2}\right)^{\frac{M}{2}-\frac{n(\boldsymbol{u})}{2}} \left|\mathbf{X}_\lambda^T\mathbf{X}_\lambda\right|^{-\frac{1}{2}} \frac{1}{\sigma_\lambda^2}\mathbb{1}_{]0,+\infty[}(\sigma_\lambda^2) \\
&\times \exp\left(-\frac{(\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda)^T(\mathbf{I}_M - \mathbf{X}_\lambda(\mathbf{X}_\lambda^T\mathbf{X}_\lambda)^{-1}\mathbf{X}_\lambda^T)(\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda)}{2\sigma_\lambda^2}\right)
\end{aligned} \tag{I.8}$$

Finalement la densité *a posteriori* jointe globale (sur les Λ feuillets du cube) s'écrit :

$$\begin{aligned} p(\boldsymbol{u}, \boldsymbol{m}, \boldsymbol{\sigma^2}|\mathbf{Y}) \propto & \prod_{\lambda=1}^{\Lambda} \Big\{ p(m_\lambda, \sigma_\lambda^2|\boldsymbol{u}, \boldsymbol{y}_\lambda) \Big\} p(\boldsymbol{u}) \\ \propto & \prod_{\lambda=1}^{\Lambda} \Big\{ \left(\frac{1}{2\pi\sigma_\lambda^2}\right)^{\frac{M-n(\boldsymbol{u})}{2}} e^{-\frac{(\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda)^T \mathbf{V}_\lambda (\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda)}{2\sigma_\lambda^2}} \left|\mathbf{X}_\lambda^T \mathbf{X}_\lambda\right|^{-\frac{1}{2}} \times \frac{1}{\sigma_\lambda^2} \mathbb{1}_{]0,+\infty[}(\sigma_\lambda^2) \Big\} \\ & \times \Gamma\left(n(\boldsymbol{u}) + 1\right) \times q^{n(\boldsymbol{u})+1} h(\boldsymbol{u}). \end{aligned} \tag{I.9}$$

avec $\mathbf{V}_\lambda = \mathbf{I}_M - \mathbf{X}_\lambda (\mathbf{X}_\lambda^T \mathbf{X}_\lambda)^{-1} \mathbf{X}_\lambda^T$.

I.2 Lois *a posteriori* conditionnelles des paramètres du bruit

De l'équation (I.9), nous pouvons déduire la loi *a posteriori* conditionnelle de σ_λ. En notant $\beta = \frac{1}{2}(\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda)^T \mathbf{V}_\lambda (\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda)$ et $\alpha = \frac{M-n(\boldsymbol{u})}{2}$, la densité $p(m_\lambda, \sigma_\lambda^2|\boldsymbol{u}, \boldsymbol{y}_\lambda)$ devient :

$$p(m_\lambda, \sigma_\lambda^2|\boldsymbol{u}, \boldsymbol{y}_\lambda) \propto \frac{\Gamma(\alpha)}{\beta^\alpha} \times \underbrace{\frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{\sigma_\lambda^2}\right)^{\alpha+1} e^{-\frac{\beta}{\sigma_\lambda^2}}}_{\sigma_\lambda^2|\boldsymbol{u}, m_\lambda \sim \mathcal{IG}(\alpha, \beta)} \left|\mathbf{X}_\lambda^T \mathbf{X}_\lambda\right|^{-\frac{1}{2}} \times \left(\frac{1}{2\pi}\right)^{\frac{M-n(\boldsymbol{u})}{2}} \mathbb{1}_{]0,+\infty[}(\sigma_\lambda^2) \tag{I.10}$$

En marginalisant l'équation (I.10) selon σ_λ, nous pouvons ensuite obtenir la densité *a posteriori* conditionnelle de m_λ :

$$\begin{aligned} p(m_\lambda|\boldsymbol{u}, \boldsymbol{y}_\lambda) \propto & \int p(m_\lambda, \sigma_\lambda^2|\boldsymbol{u}, \boldsymbol{y}_\lambda) d\sigma_\lambda^2 \\ \propto & \frac{\Gamma(\alpha)}{\beta^\alpha} \times \int \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{\sigma_\lambda^2}\right)^{\alpha+1} e^{-\frac{\beta}{\sigma_\lambda^2}} \times \mathbb{1}_{]0,+\infty[}(\sigma_\lambda^2)\, d\sigma_\lambda^2 \left(\frac{1}{2\pi}\right)^{\frac{M-n(\boldsymbol{u})}{2}} \left|\mathbf{X}_\lambda^T \mathbf{X}_\lambda\right|^{-\frac{1}{2}} \\ \propto & \frac{\Gamma(\alpha)}{\beta^\alpha} \times \left(\frac{1}{2\pi}\right)^{\frac{M-n(\boldsymbol{u})}{2}} \times \left|\mathbf{X}_\lambda^T \mathbf{X}_\lambda\right|^{-\frac{1}{2}} \end{aligned} \tag{I.11}$$

Afin de faire apparaître la densité *a posteriori* conditionnelle de m_λ, il faut effectuer quelques transformations sur l'expression de β :

$$\begin{aligned} \beta &= \frac{1}{2}(\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda)^T \mathbf{V}_\lambda (\boldsymbol{y}_\lambda - \mathbf{1}m_\lambda) && \text{(I.12)} \\ &= \frac{1}{2}(\boldsymbol{y}_\lambda^T \mathbf{V}_\lambda \boldsymbol{y}_\lambda - 2m_\lambda \mathbf{1}^T \mathbf{V}_\lambda \boldsymbol{y}_\lambda + m_\lambda^2 \mathbf{1}^T \mathbf{V}_\lambda \mathbf{1}) && \text{(I.13)} \\ &= \frac{1}{2}(\boldsymbol{y}_\lambda^T \mathbf{V}_\lambda \boldsymbol{y}_\lambda + (m_\lambda - \tilde{m}_\lambda) \mathbf{1}^T \mathbf{V}_\lambda \mathbf{1} (m_\lambda - \tilde{m}_\lambda) - \tilde{m}_\lambda \mathbf{1}^T \mathbf{V}_\lambda \mathbf{1} \tilde{m}_\lambda) && \text{(I.14)} \end{aligned}$$

avec

$$\begin{aligned} \mathbf{V}_\lambda &= \mathbf{I}_M - \mathbf{X}_\lambda (\mathbf{X}_\lambda^T \mathbf{X}_\lambda)^{-1} \mathbf{X}_\lambda^T, \\ \delta_\lambda^2 &= (\mathbf{1}^T \mathbf{V}_\lambda \mathbf{1})^{-1} \\ \tilde{m}_\lambda &= \delta_\lambda^2 \mathbf{1}^T \mathbf{V}_\lambda \boldsymbol{y}_\lambda, \end{aligned}$$

L'équation (I.14) peut se réécrire :

$$\beta = \frac{1}{2}(\boldsymbol{y}_\lambda^T \mathbf{V}_\lambda \boldsymbol{y}_\lambda + (m_\lambda - \tilde{m}_\lambda)\mathbf{1}^T\mathbf{V}_\lambda\mathbf{1}(m_\lambda - \tilde{m}_\lambda) - \delta_\lambda^2 \mathbf{1}^T\mathbf{V}_\lambda\boldsymbol{y}_\lambda \cancel{\mathbf{1}^T\mathbf{V}_\lambda\mathbf{1}\delta_\lambda^2}\mathbf{1}^T\mathbf{V}_\lambda\boldsymbol{y}_\lambda) \tag{I.15}$$

$$= \frac{1}{2}(\boldsymbol{y}_\lambda^T \mathbf{V}_\lambda \boldsymbol{y}_\lambda + (m_\lambda - \tilde{m}_\lambda)\frac{1}{\delta_\lambda^2}(m_\lambda - \tilde{m}_\lambda) - \delta_\lambda^2(\mathbf{1}^T\mathbf{V}_\lambda\boldsymbol{y}_\lambda)^2) \tag{I.16}$$

$$= \frac{1}{2\delta_\lambda^2}\left(\delta_\lambda^2\left(\boldsymbol{y}_\lambda^T \mathbf{V}_\lambda \boldsymbol{y}_\lambda - \delta_\lambda^2(\mathbf{1}^T\mathbf{V}_\lambda\boldsymbol{y}_\lambda)^2\right) + (m_\lambda - \tilde{m}_\lambda)^2\right) \tag{I.17}$$

$$= \frac{\nu s_\lambda^2}{2\delta_\lambda^2}\left(1 + \frac{(m_\lambda - \tilde{m}_\lambda)^2}{\nu s_\lambda^2}\right) \tag{I.18}$$

en posant

$$\begin{aligned}\nu &= M - 1 - n(\boldsymbol{u}),\\ s_\lambda^2 &= \nu^{-1}\delta_\lambda^2\left[\boldsymbol{y}_\lambda^T \mathbf{V}_\lambda \boldsymbol{y}_\lambda - \delta_\lambda^2(\mathbf{1}^T\mathbf{V}_\lambda\boldsymbol{y}_\lambda)^2\right].\end{aligned}$$

L'introduction du facteur ν dans l'expression de s_λ^2 n'est pas anodine puisque $\alpha = \frac{\nu+1}{2}$, ce qui permet d'écrire :

$$\begin{aligned}p(m_\lambda|\boldsymbol{u},\boldsymbol{y}_\lambda) &\propto \frac{\Gamma(\alpha)}{\beta^\alpha} \times \left(\frac{1}{2\pi}\right)^{\frac{M-n(\boldsymbol{u})}{2}} \times \left|\mathbf{X}_\lambda^T\mathbf{X}_\lambda\right|^{-\frac{1}{2}}\\ &\propto \Gamma\left(\frac{M-n(\boldsymbol{u})}{2}\right)\left(\frac{\nu s_\lambda^2}{2\delta_\lambda^2}\left(1 + \frac{(m_\lambda - \tilde{m}_\lambda)^2}{\nu s_\lambda^2}\right)\right)^{\frac{\nu+1}{2}} \times \left(\frac{1}{2\pi}\right)^{\frac{M-n(\boldsymbol{u})}{2}} \times \left|\mathbf{X}_\lambda^T\mathbf{X}_\lambda\right|^{-\frac{1}{2}}\\ &\propto \left(1 + \frac{1}{\nu}\left(\frac{m_\lambda - \tilde{m}_\lambda}{s_\lambda}\right)^2\right)^{\frac{\nu+1}{2}} \times \left(\frac{\nu s_\lambda^2}{2\delta_\lambda^2}\right)^{\frac{\nu+1}{2}}\Gamma\left(\frac{\nu+1}{2}\right) \times \left(\frac{1}{2\pi}\right)^{\frac{M-n(\boldsymbol{u})}{2}} \times \left|\mathbf{X}_\lambda^T\mathbf{X}_\lambda\right|^{-\frac{1}{2}}\\ &\propto \left(1 + \frac{1}{\nu}\left(\frac{m_\lambda - \tilde{m}_\lambda}{s_\lambda}\right)^2\right)^{\frac{\nu+1}{2}}\end{aligned} \tag{I.19}$$

Finalement, la loi conditionnelle *a posteriori* de m_λ une loi de Student à ν degrés de liberté et où $\tilde{m}_\lambda$ est le paramètre de localisation et s_λ est le paramètre d'échelle.

Bibliographie

E. Arias-Castro, E. J. Candès, Y. Plan, et al. Global testing under sparse alternatives : ANOVA, multiple comparisons and the higher criticism. *The Annals of Statistics*, 39(5) :2533–2556, 2011.

R. Bacon, J. Brinchmann, J. Richard, T. Contini, A. Drake, M. Franx, S. Tacchella, J. Vernet, L. Wisotzki, J. Blaizot, et al. The MUSE 3D view of the Hubble Deep Field South. *Astronomy and Astrophysics*, 575, 2015.

A. J. Baddeley and M. N. M. Van Lieshout. Stochastic geometry models in high-level vision. *Journal of Applied Statistics*, 20(5-6) :231–256, 1993.

R. F. Barber and E. Candes. Controlling the false discovery rate via knockoffs. *arXiv preprint arXiv :1404.5609*, 2014.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate : a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

E. Bertin. Sextractor user's manual. https ://www.astromatic.net/pubsvn/software/sextractor/trunk/doc/sextractor.pdf.

E. Bertin and S. Arnouts. SExtractor : Software for source extraction. *Astronomy and Astrophysics Supplement 317*, 117 :393–404, June 1996.

J. W. Boardman. Automating spectral unmixing of aviris data using convex geometry concepts. 1993.

S. Bourguignon, D. Mary, and É. Slezak. Restoration of astrophysical spectra with sparsity constraints : Models and algorithms. *Selected Topics in Signal Processing, IEEE Journal of*, 5 (5) :1002–1013, 2011.

S. Cantalupo. CubEx. In *Busy week, MUSE Consortium*, 2014.

H. Carfantan. Habilitation à diriger les recherches : Modèles estimateurs et algorithmes pour quelques problèmes inverses de traitement du signal et d'images en sciences de l'univers, 2014. http ://userpages.irap.omp.eu/ hcarfantan/Carfantan_HdR-memoire.pdf.

S. Casertano, D. de Mello, M. Dickinson, H. C. Ferguson, A. S. Fruchter, R. A. Gonzalez-Lopezlira, I. Heyer, R. N. Hook, Z. Levay, R. A. Lucas, et al. WFPC2 Observations of the Hubble Deep Field South. *The Astronomical Journal*, 120(6) :2747, 2000.

G. Celeux, J.-M. Marin, and C. Robert. Sélection bayésienne de variables en régression linéaire. *Journal de la société française de statistique*, 147(1) :59–79, 2006.

G. Celeux, M. El Anbari, J.-M. Marin, C. P. Robert, et al. Regularization in regression : comparing bayesian and frequentist methods in a poorly informative situation. *Bayesian Analysis*, 7 (2) :477–502, 2012.

F. Chatelain, X. Descombes, and J. Zerubia. *Energy minimization methods in Computer Vision and Pattern Recognition*, chapter Parameter Estimation for Marked Point Processes. Application to Object Extraction from Remote Sensing Images, pages 221–234. Daniel Cremers, Yuri Boykov, Andrew Blake, and Frank R. Schmidt, 2009.

F. Chatelain, A. Costard, and O. Michel. A Bayesian marked point process for object detection. Application to MUSE hyperspectral data. In *Proc. IEEE International Conference Acoustics, Speech and Signal Processing (ICASSP)*, 2011.

D. Cheng and A. Schwartzman. Multiple testing of local maxima for detection of peaks in random fields. *arXiv preprint arXiv :1405.1400*, 2014.

S. N. Chiu, D. Stoyan, W. S. Kendall, and J. Mecke. *Stochastic geometry and its applications*. John Wiley & Sons, 2013.

G. De Vaucouleurs. General physical properties of external galaxies. In *Astrophysik IV : Sternsysteme/Astrophysics IV : Stellar Systems*, pages 311–372. Springer, 1959.

S. Descamps, X. Descombes, and J. Zerubia. Automatic flamingo detection using a multiple birth and death process. In *Proc. IEEE International Conference Acoustics, Speech and Signal Processing (ICASSP)*, 2008.

X. Descombes. *Stochastic geometry for image analysis*. Wiley-ISTE, 2011. URL https://hal.inria.fr/hal-00793677.

X. Descombes, R. Minlos, and E. Zhizhina. Object extraction using a stochastic birth-and-death dynamics in continuum. *Journal of Mathematical Imaging and Vision*, 33(3) :347–359, 2009.

P. J. Diggle and R. K. Milne. Negative binomial quadrat counts and point processes. *Scandinavian Journal of Statistics*, 10(4) :257–267, 1983.

D. Donoho and J. Jin. Higher Criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*, pages 962–994, 2004.

B. Efron. *Large-scale inference : empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2010.

B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

H. C. Ferguson, M. Dickinson, and R. Williams. The Hubble Deep Fields. *Annual Review of Astronomy and Astrophysics*, 38 :667–715, 2000.

A. Fruchter et al. Hst multidrizzle handbook. *HST MultiDrizzle, HST Data Handbooks*, 1, 2009.

T. Garel. *Modélisation de l'émission Lyman-alpha dans les galaxies à grand décalage spectral et simulations cosmologiques*. PhD thesis, Université Claude Bernard-Lyon I, 2011.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distribution and Bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pages 721–741, 1984.

C. R. Genovese, N. A. Lazar, and T. Nichols. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, 15(4) :870–878, 2002.

C. J. Geyer and J. Møller. Simulation procedures and likelihood inference for spatial point processes. *Scandinavian Journal of Statistics*, 21(4) :pp. 359–373, 1994.

P. Green. Reversible Jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 52 :711–732, 1995.

S. B. Hadj, F. Chatelain, X. Descombes, and J. Zerubia. Approche non supervisée par processus ponctuels marqués pour l'extraction d'objets à partir d'images aériennes et satellitaires. *Revue Française de Photogrammétrie et de Télédétection*, 194 :2–15, 2010.

P. Hall, J. Jin, et al. Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics*, 38(3) :1686–1732, 2010.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57 :97–109, 1970.

C. Herenz. LSDcat. 2014.

D. W. Hogg, M. A. Pahre, K. L. Adelberger, R. Blandford, J. G. Cohen, T. Gautier, T. Jarrett, G. Neugebauer, and C. C. Steidel. Caltech faint galaxy redshift survey. ix. source detection and photometry in the hubble deep field region. *The Astrophysical Journal Supplement Series*, 127(1) :1, 2000.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

B. Holwerda. Sextractor for dummies. http ://mensa.ast.uct.ac.za/∼holwerda/SE/Manual_files/Guide2source_extractor.pdf.

A. M. Hopkins, C. J. Miller, A. Connolly, C. Genovese, R. C. Nichol, and L. Wasserman. A new source detection algorithm using the false-discovery rate. *The Astronomical Journal*, 123(2) : 1086–1094, 2002.

T.-O. Husser. *3D-Spectroscopy of Dense Stellar Populations*. Universitätsverlag Göttingen, 2012.

A. Jarno, R. Bacon, P. Ferruit, and A. Pécontal-Rousset. Numerical simulation of the VLT/MUSE instrument. In *SPIE Astronomical Telescopes+ Instrumentation*, pages 701710–701710. International Society for Optics and Photonics, 2008.

H. Jeffreys. *Theory of Probability (third edition)*. Oxford University Press, 1961.

R. Kass and L. Wasserman. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91 :1343–1370, 1996.

E. J. Kelly. An adaptive detection algorithm. *Aerospace and Electronic Systems, IEEE Transactions on*, (2) :115–127, 1986.

W. S. Kendall and J. Møller. Perfect simulation using dominating processes on ordered spaces, with application to locally stable point processes. *Advances in Applied Probability*, pages 844–865, 2000.

B. Keresztes, O. Lavialle, and M. Borda. Seismic fault detection using marked point processes. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 565–568, Nov 2009.

F. Laurent, F. Henault, E. Renault, R. Bacon, and J.-P. Dubois. Design of an integral field unit for muse, and results from prototyping. *Publications of the Astronomical Society of the Pacific*, 118(849) :pp. 1564–1573, 2006.

F. Laurent, E. Renault, H. Anwand, D. Boudon, P. Caillier, J. Kosmalski, M. Loupias, H. Nicklas, W. Seifert, Y. Salaun, et al. Muse field splitter unit : fan-shaped separator for 24 integral field units. In *SPIE Astronomical Telescopes+ Instrumentation*, pages 91511U–91511U. International Society for Optics and Photonics, 2014.

R. Lutz. An algorithm for the real time analysis of digitised images. *The Computer Journal*, 23 (3) :262–269, 1980.

D. Manolakis and G. A. Shaw. Detection algorithms for hyperspectral imaging application. *IEEE Signal Processing Magazine*, January 2002.

D. Manolakis, D. Marden, and G. A. Shaw. Hyperspectral image processing for automatic target detection applications. *Lincoln Laboratory Journal*, 14, 2003.

D. Manolakis, E. Truslow, M. Pieper, T. Cooley, and M. Brueggeman. Detection algorithms in hyperspectral imaging systems : An overview of practical algorithms. *Signal Processing Magazine, IEEE*, 31(1) :24–33, Jan 2014.

D. Mary, A. Ferrari, and S. Paris. Detection of astrophysical sources in hyperspectral data. applications to the muse instrument. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 6031–6035, Oct 2014.

C. Meillier, F. Chatelain, O. Michel, and H. Ayasso. Non-parametric bayesian framework for detection of object configurations with large intensity dynamics in highly noisy hyperspectral data. In *Proc. IEEE International Conference Acoustics, Speech and Signal Processing (ICASSP)*, pages 1886–1890, 2014. doi : 10.1109/ICASSP.2014.6853926.

C. Meillier, F. Chatelain, O. Michel, and H. Ayasso. Nonparametric bayesian extraction of object configurations in massive data. *Signal Processing, IEEE Transactions on*, 63(8) :1911–1924, 2015a. ISSN 1053-587X. doi : 10.1109/TSP.2015.2403268.

C. Meillier, F. Chatelain, O. Michel, and H. Ayasso. Error control for the detection of rare and weak signatures in massive data. In *Proc. 23rd European Signal Processing Conference (EUSIPCO 2015)*, 2015b.

C. Meillier, F. Chatelain, O. Michel, and H. Ayasso. Contrôle des erreurs pour la détection d'événements rares et faibles dans des champs de données massifs. In *GRETSI 2015*, 2015c.

C. Meillier, F. Chatelain, O. Michel, R. Bacon, L. Piqueras, R. Bacher, and H. Ayasso. Selfi : an object based, bayesian method for faint emission line source detection in muse deep field datacubes. *Astronomy and Astrophysics*, 2016.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, 21(6) :1087–1092, 1953.

C. J. Miller, C. Genovese, R. C. Nichol, L. Wasserman, A. Connolly, D. Reichart, A. Hopkins, J. Schneider, and A. Moore. Controlling the false-discovery rate in astrophysical data analysis. *The Astronomical Journal*, 122(6) :3492, 2001.

A. Moffat. A theoretical investigation of focal stellar images in the photographic emulsion and application to photographic photometry. *Astronomy and Astrophysics*, 3 :455, 1969.

R. Molina and B. D. Ripley. Using spatial models as priors in astronomical image analysis. *Journal of Applied Statistics*, 20(5-6) :281–298, 1993.

J. M. Nascimento and J. M. Bioucas Dias. Vertex component analysis : A fast algorithm to unmix hyperspectral data. *Geoscience and Remote Sensing, IEEE Transactions on*, 43(4) : 898–910, 2005.

N. Nasrabadi. Hyperspectral target detection : An overview of current and future challenges. *Signal Processing Magazine, IEEE*, 31(1) :34–44, 2014.

M. Ortner. *Marked point processes for extracting buildings from digital elevation models*. Theses, Université Nice Sophia Antipolis, Oct. 2004. URL https://tel.archives-ouvertes.fr/tel-00189803.

S. Paris. *Sparsity-based detection strategies for faint signals in noise : application to astrophysical hyperspectral data*. Theses, Université Nice Sophia Antipolis ; Università degli studi La Sapienza (Rome), Oct. 2013. URL https://tel.archives-ouvertes.fr/tel-00933827.

S. Paris, D. Mary, A. Ferrari, and S. Bourguigon. Sparsity-based composite detection tests. application to astrophysical hyperspectral data. In *Proc. 19th European Signal Processing Conference (EUSIPCO 2011)*, pages 1909–1913, 2011.

S. Paris, D. Mary, and A. Ferrari. Detection tests using sparse models, with application to hyperspectral data. *IEEE Trans. on signal processing*, 61(5-8) :1481–1494, 2013a.

S. Paris, R. Suleiman, D. Mary, and A. Ferrari. Constrained likelihood ratios for detecting sparse signals in highly noisy 3D data. In *Proc. IEEE International Conference Acoustics, Speech and Signal Processing (ICASSP)*, 2013b.

C. Y. Peng, L. C. Ho, C. D. Impey, and H.-W. Rix. Detailed structural decomposition of galaxy images. *The Astronomical Journal*, 124(1), 2002.

C. Y. Peng, L. C. Ho, C. D. Impey, and H.-W. Rix. Detailed decomposition of galaxy images. ii. beyond axisymmetric models. *The Astronomical Journal*, 139(6) :2097, 2010.

B. Perret, V. Mazet, C. Collet, and E. SIEZAK. Décomposition d'images multispectrales de galaxies au moyen d'algorithmes de monte carlo par chaines de markov. In *XXIIe colloque GRETSI (traitement du signal et des images), Dijon (FRA), 8-11 septembre 2009*. GRETSI, Groupe d ?Etudes du Traitement du Signal et des Images, 2009.

A. Popping, R. Jurek, T. Westmeier, P. Serra, L. Flöer, M. Meyer, and B. Koribalski. Comparison of Potential ASKAP Hi Survey Source Finders. *Publications of the Astronomical Society of Australia*, 29 :318–339, 2012.

J. Richard, V. Patricio, J. Martinez, R. Bacon, B. Clément, P. Weilbacher, K. Soto, L. Wisotzki, J. Vernet, R. Pello, et al. Muse observations of the lensing cluster smacsj2031. 8-4036 : new constraints on the mass distribution in the cluster core. *Monthly Notices of the Royal Astronomical Society : Letters*, 446(1) :L16–L20, 2015.

C. Robert. *Le choix bayésien : Principes et pratique*. Springer Science & Business Media, 2006.

A. Schwartzman, Y. Gavrilov, and R. J. Adler. Multiple testing of local maxima for detection of peaks in 1d. *Annals of statistics*, 39(6) :3290, 2011.

P. Serra, R. Jurek, and L. Flöer. Using negative detections to estimate source-finder reliability. *Publications of the Astronomical Society of Australia*, 29(03) :296–300, 2012a.

P. Serra, T. Oosterloo, R. Morganti, K. Alatalo, L. Blitz, M. Bois, F. Bournaud, M. Bureau, M. Cappellari, A. F. Crocker, R. L. Davies, T. A. Davis, P. T. de Zeeuw, P.-A. Duc, E. Emsellem, S. Khochfar, D. Krajnovi ?, H. Kuntschner, P.-Y. Lablanche, R. M. McDermid, T. Naab, M. Sarzi, N. Scott, S. C. Trager, A.-M. Weijmans, and L. M. Young. The ATLAS3D project ? XIII. Mass and morphology of H ?i in early-type galaxies as a function of environment. *Monthly Notices of the Royal Astronomical Society*, 422(3) :1835–1862, 2012b.

P. Serra, T. Westmeier, N. Giese, R. Jurek, L. Flöer, A. Popping, B. Winkel, T. van der Hulst, M. Meyer, B. S. Koribalski, L. Staveley-Smith, and H. Courtois. SoFiA : a flexible source finder for 3D spectral line data. *Monthly Notices of the Royal Astronomical Society*, 448(2) : 1922–1929, 2015.

D. Serre, E. Villeneuve, H. Carfantan, L. Jolissaint, V. Mazet, S. Bourguignon, and A. Jarno. Modeling the spatial PSF at the VLT focal plane for MUSE WFM data analysis purpose. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 7736, pages 773649–1 – 773649–12, 2010.

J. L. Sersic. Influence of the atmospheric and instrumental dispersion on the brightness distribution in a galaxy. *Bulletin of the Astronomical Association of Argentina*, pages 41–43, 1963.

Y.-S. Shen, T.-H. Chan, S. Bourguignon, and C.-Y. Chi. Spatial-spectral unmixing of hyperspectral data for detection and analysis of astrophysical sources with the muse instrument. In *IEEE WHISPERS*, 2012.

L. Simard, C. N. A. Willmer, N. P. Vogt, V. L. Sarajedini, A. C. Phillips, B. J. Weiner, D. C. Koo, M. Im, G. D. Illingworth, and S. M. Faber. The deep groth strip survey. ii. hubble space telescope structural parameters of galaxies in the groth strip. *The Astrophysical Journal Supplement Series*, 142(1) :1, 2002.

M. Smith and R. Kohn. Nonparametric regression using bayesian variable selection. *Journal of Econometrics*, 75(2) :317–343, 1996.

R. Stoica, X. Descombes, and J. Zerubia. A gibbs point process for road extraction from remotely sensed images. *International Journal of Computer Vision*, 57(2) :121–136, 2004.

J. D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 64(3) :479–498, 2002.

R. Suleiman, D. Mary, and A. Ferrari. Minimax sparse detection based on one-class classifiers. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 5553–5557. IEEE, 2013.

Y. Taniguchi, N. Scoville, T. Murayama, D. Sanders, B. Mobasher, H. Aussel, P. Capak, M. Ajiki, S. Miyazaki, Y. Komiyama, et al. The cosmic evolution survey (cosmos) : Subaru observations of the hst cosmos field. *The Astrophysical Journal Supplement Series*, 172(1) :9, 2007.

C. Tapken, I. Appenzeller, S. Noll, S. Richling, J. Heidt, E. Meinköhn, and D. Mehlert. Lyman-α emission in high-redshift galaxies. *Astronomy & Astrophysics*, 467(1) :63–72, 2007.

O. Tournaire, N. Paparoditis, and F. Lafarge. Rectangular road marking detection with marked point processes. In *Proc. conference on Photogrammetric Image Analysis*, 2007.

I. Trujillo, J. Aguerri, J. Cepa, and C. Gutiérrez. The effects of seeing on Sérsic profiles–II. The Moffat PSF. *Monthly Notices of the Royal Astronomical Society*, 328(3) :977–985, 2001.

J. W. Tukey. T13 N : The Higher Criticism. *Course notes, Princeton University*, 1976.

S. Valero, P. Salembier, and J. Chanussot. New hyperspectral data representation using binary partition tree. In *Geoscience and Remote Sensing Symposium (IGARSS), 2010 IEEE International*, pages 80–83. IEEE, 2010.

S. Valero, P. Salembier, J. Chanussot, and C. M. Cuadras. Improved binary partition tree construction for hyperspectral images : application to object detection. In *Proc. IGARSS*, pages 1–4, 2011.

M. N. M. Van Lieshout. *Markov Point Processes and Their Applications*. London, Imperial College Press, 2000.

E. Villeneuve. *Déconvolution de données hyperspectrales pour l'instrument MUSE du VLT*. PhD thesis, Université de Toulouse, Université Toulouse III-Paul Sabatier, 2012.

E. Villeneuve, H. Carfantan, and D. Serre. Psf estimation of hyperspectral data acquisition system for ground-based astrophysical observations. In *Hyperspectral Image and Signal Processing : Evolution in Remote Sensing (WHISPERS), 2011 3rd Workshop on*, pages 1–4. IEEE, 2011.

P. M. Weilbacher, O. Streicher, T. Urrutia, A. Jarno, A. Pécontal-Rousset, R. Bacon, and P. Bohm. Design and capabilities of the muse data reduction software and pipeline, 2012.

T. Westmeier, A. Popping, and P. Serra. Basic testing of the duchamp source finder. *Publications of the Astronomical Society of Australia*, 29(3) :276–295, 2012.

M. T. Whiting. DUCHAMP : a 3D source finder for spectral-line data. *Monthly Notices of the Royal Astronomical Society*, 421(4) :3242–3256, 2012.

M. E. Winter. N-findr : an algorithm for fast autonomous spectral end-member determination in hyperspectral data. In *SPIE's International Symposium on Optical Science, Engineering, and Instrumentation*, pages 266–275. International Society for Optics and Photonics, 1999.

A. Zellner. On assessing prior distributions and bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques : Essays in Honor of Bruno De Finetti*, 6 :233–243, 1986.

Résumé — Dans cette thèse, nous nous sommes intéressés à la détection de galaxies lointaines dans les données hyperspectrales MUSE. Ces galaxies, en particulier, sont difficiles à observer, elles sont spatialement peu étendues du fait de leur distance, leur spectre est composé d'une seule raie d'émission dont la position est inconnue et dépend de la distance de la galaxie, et elles présentent un rapport signal-à-bruit très faible. Ces galaxies lointaines peuvent être considérées comme des sources quasi-ponctuelles dans les trois dimensions du cube. Il existe peu de méthodes dans la littérature qui permettent de détecter des sources dans des données en trois dimensions. L'approche proposée dans cette thèse repose sur la modélisation de la configuration de galaxies par un processus ponctuel marqué. Ceci consiste à représenter la position des galaxies comme une configuration de points auxquels nous ajoutons des caractéristiques géométriques, spectrales, etc, qui transforment un point en objet. Cette approche présente l'avantage d'avoir une représentation mathématique proche du phénomène physique et permet de s'affranchir des approches pixelliques qui sont pénalisées par les dimensions conséquentes des données ($300 \times 300 \times 3600$ pixels). La détection des galaxies et l'estimation de leurs caractéristiques spatiales, spectrales ou d'intensité sont réalisées dans un cadre entièrement bayésien, ce qui conduit à un algorithme générique et robuste, où tous les paramètres sont estimés sur la base des seules données observées, la détection des objets d'intérêt étant effectuée conjointement. La dimension des données et la difficulté du problème de détection nous ont conduit à envisager une phase de prétraitement des données visant à définir des zones de recherche dans le cube. Des approches de type tests multiples permettent de construire des cartes de proposition des objets. La détection bayésienne est guidée par ces cartes de pré-détection (définition de la fonction d'intensité du processus ponctuel marqué), la proposition des objets est réalisée sur les pixels sélectionnés sur ces cartes. La qualité de la détection peut être caractérisée par un critère de contrôle des erreurs. L'ensemble des traitements développés au cours de cette thèse a été validé sur des données synthétiques, et appliqué ensuite à un jeu de données réelles acquises par MUSE suite à sa mise en service en 2014. L'analyse de la détection obtenue est présentée dans le manuscrit.

Mots clés : Détection, estimation, processus ponctuels marqués, tests multiples, hyperspectral.

Abstract — Detecting the faintest galaxies in the hyperspectral MUSE data is particularly challenging because they have a small spatial extension, a very sparse spectrum that contains only one narrow emission line, which position in the spectral range is unknown. Moreover, their signal-to-noise ratio are very low. These galaxies are modeled as quasi point sources in the three dimensions of the data cube. We propose a method for the detection of a galaxy configuration based on a marked point process in a nonparametric Bayesian framework. A galaxy is modeled by a point (its position in the spatial domain), and marks (geometrical, spectral features) are added to transform a point into an object. These processes yield a natural sparse representation of massive data ($300 \times 300 \times 3600$ pixels). The fully Bayesian framework leads to a general and robust algorithm where the parameters of the objects are estimated in a fully data-driven way. Preprocessing strategies are drawn to tackle the massive dimensions of the data and the complexity of the detection problem, they allow to reduce the exploration of the data to areas that probably contain sources. Multiple testing approaches have been proposed to build proposition map. This map is also used to define the intensity of the point process, *i.e.* it describes the probability density function of the point process. It also gives a global error control criterion for the detection. The performance of the proposed algorithm is illustrated on synthetic data and real hyperspectral data acquired by the MUSE instrument for young galaxy detection.

Keywords : Detection, estimation, marked point process, multiple testing problem, hyperspectral.

Gipsa-lab, 11 rue des mathématiques
38400 Saint Martin d'Hères